

# Learning the Second-Moment Matrix of a Smooth Function From Point Samples

Armin Eftekhari, Michael B. Wakin, Ping Li, Paul G. Constantine, and Rachel A. Ward\*

January 30, 2017

## Abstract

Consider an open set  $\mathbb{D} \subseteq \mathbb{R}^n$ , equipped with a probability measure  $\mu$ . An important characteristic of a smooth function  $f : \mathbb{D} \rightarrow \mathbb{R}$  is its *second-moment matrix*  $\Sigma_\mu := \int \nabla f(x)(\nabla f(x))^* \mu(dx) \in \mathbb{R}^{n \times n}$ , where  $\nabla f(x) \in \mathbb{R}^n$  is the gradient of  $f(\cdot)$  at  $x \in \mathbb{D}$ . For instance, the span of the leading  $r$  eigenvectors of  $\Sigma_\mu$  forms an *active subspace* of  $f(\cdot)$ , thereby extending the concept of *principal component analysis* to the problem of *ridge approximation*. In this work, we propose a simple algorithm for estimating  $\Sigma_\mu$  from point values of  $f(\cdot)$  *without* imposing any structural assumptions on  $f(\cdot)$ . Theoretical guarantees for this algorithm are provided with the aid of the same technical tools that have proved valuable in the context of covariance matrix estimation from partial measurements.

## 1 Introduction

Central to approximation theory, machine learning, and computational sciences in general is the task of extrapolating a function given its finitely many point samples. More concretely, consider an open set  $\mathbb{D} \subseteq \mathbb{R}^n$ , equipped with subspace Borel  $\sigma$ -algebra and probability measure  $\mu$ . The objective is to *learn* (approximate) a smooth function  $f : \mathbb{D} \rightarrow \mathbb{R}$  from the query points

$$\{x_i\}_{i=1}^N \subset \mathbb{D},$$

and evaluation of  $f(\cdot)$  at these points [1, 2, 3, 4].

An important quantity in this context is the *second-moment matrix* of  $f(\cdot)$  with respect to the measure  $\mu$ , defined as

$$\Sigma_\mu := \mathbb{E}_x [\nabla f(x) \cdot (\nabla f(x))^*] = \int_{\mathbb{D}} \nabla f(x) \cdot (\nabla f(x))^* \mu(dx) \in \mathbb{R}^{n \times n}, \quad (1)$$

where  $\nabla f(x) \in \mathbb{R}^n$  is the gradient of  $f(\cdot)$  at  $x \in \mathbb{D}$  and the superscript  $*$  denotes vector and matrix transpose.<sup>1</sup> Roughly speaking, in this matrix,  $\Sigma_\mu[i, j]$ , the  $[i, j]$ th entry of  $\Sigma_\mu$ , measures the correlation between the  $i$ th and  $j$ th partial derivatives of  $f(\cdot)$ . Note that  $\Sigma_\mu$  captures key information about how  $f(\cdot)$  changes along different directions. Indeed, for an arbitrary vector  $v \in \mathbb{R}^n$  with  $\|v\|_2 = 1$ , the *directional derivative* of  $f(\cdot)$  at  $x \in \mathbb{D}$  and along  $v$  is  $v^* \nabla f(x)$ , and the “energy” of the directional derivative of  $f(\cdot)$  along  $v$  and with respect to  $\mu$  is  $v^* \Sigma_\mu v$ . Furthermore, in ridge approximation, the leading  $r$  eigenvectors of  $\Sigma_\mu$  span an  $r$ -dimensional *active subspace* of  $f(\cdot)$  with respect to the measure  $\mu$  [5]. If  $U_{\mu,r} \in \mathbb{R}^{n \times r}$  denotes an orthonormal basis for this active subspace, then it might be possible to reliably approximate  $f(x)$  with  $h(U_{\mu,r}^* x)$  for all  $x \in \mathbb{D}$  and for some smooth function  $h : \mathbb{R}^r \rightarrow \mathbb{R}$ . Beyond approximation theory, the significance of second-moment matrices (and related concepts) across a number of other disciplines is discussed in Section 4.

With this introduction, let us now state the main objective of this paper (which we will make precise later in Section 2).

\*AE is with the Alan Turing Institute in London. MBW is with the Electrical Engineering and Computer Science department at the Colorado School of Mines. PL is with the Statistics and Computer Science departments at Rutgers University. PGC is with the Applied Mathematics and Statistics department at the Colorado School of Mines. RAW is with the Mathematics department at the University of Texas, Austin, and the Institute for Computational Engineering and Sciences.

<sup>1</sup>As suggested above, we will often suppress the dependence on  $f(\cdot)$  in our notation for the sake of brevity.

**Objective:** Design query points  $\{x_i\}_{i=1}^N$  and learn the second-moment matrix of  $f(\cdot)$  (with respect to measure  $\mu$ ) from  $\{x_i, f(x_i)\}_{i=1}^N$ .

We must emphasize that we impose *no structural assumptions* on the second-moment matrix (such as being low rank or sparse), a point that we shall revisit later in Sections 1.2 and 4. Our approach to this problem, alongside the results, are summarized next with minimal details for better accessibility. A rigorous account of the problem and our approach is then presented in Sections 2 and 3.

## 1.1 Approach

Consider  $N$  random points drawn independently from  $\mu$  and stored as the columns of  $X \in \mathbb{R}^{n \times N}$ . It is then easy to verify that

$$\dot{\Sigma}_X := \frac{1}{N} \sum_{x \in X} \nabla f(x) \cdot \nabla f(x)^* \quad (2)$$

is an unbiased estimator of  $\Sigma_\mu$  in (1).<sup>2</sup> In fact, a standard large deviation analysis reveals that  $\|\dot{\Sigma}_X - \Sigma_\mu\| \propto \frac{1}{\sqrt{N}}$  with overwhelming probability and for any matrix norm  $\|\cdot\|$ .

Since only the point values of  $f(\cdot)$  are at our disposal (rather than its gradient), it is not possible to directly calculate  $\dot{\Sigma}_X$  and one might resort to using forward Euler methods, as we sketch here and formalize in Section 2. For a sufficiently small  $\epsilon > 0$  and arbitrary  $x$ , let  $\mathbb{B}_{x,\epsilon}$  denote the Euclidean ball of radius  $\epsilon$  about  $x$ , and set

$$\mathbb{B}_{X,\epsilon} = \bigcup_{x \in X} \mathbb{B}_{x,\epsilon}.$$

Let also  $\mu_{X,\epsilon}$  be the conditional probability measure on  $\mathbb{B}_{X,\epsilon}$  induced by  $\mu$ . Consider  $N_{X,\epsilon}$  random points drawn independently from  $\mu_{X,\epsilon}$  and stored as the columns of  $Y_{X,\epsilon} \in \mathbb{R}^{n \times N_{X,\epsilon}}$ . Then partition  $Y_{X,\epsilon}$  according to  $X$  by setting  $Y_{x,\epsilon} = Y_{X,\epsilon} \cap \mathbb{B}_{x,\epsilon}$ , so that  $Y_{x,\epsilon} \in \mathbb{R}^{n \times N_{x,\epsilon}}$  contains all  $\epsilon$ -neighbors of  $x$  in  $Y_{X,\epsilon}$ . This setup is illustrated in Figure 1.

For every  $x \in X$ , consider  $\dot{\nabla}_{Y_{x,\epsilon}} f(x) \in \mathbb{R}^n$  as an estimate of the true gradient  $\nabla f(x)$ , where

$$\dot{\nabla}_{Y_{x,\epsilon}} f(x) := \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} \frac{f(y) - f(x)}{\|y - x\|_2} \cdot \frac{y - x}{\|y - x\|_2}, \quad (3)$$

and the scaling with  $n$  will be shortly justified. Then, we may naturally consider  $\dot{\Sigma}_{X,Y_{X,\epsilon}} \in \mathbb{R}^{n \times n}$  as an estimate of  $\dot{\Sigma}_X$  (see (2)) and in turn an estimate of  $\Sigma_\mu$  (see (1)), where

$$\dot{\Sigma}_{X,Y_{X,\epsilon}} := \frac{1}{N} \sum_{x \in X} \dot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \dot{\nabla}_{Y_{x,\epsilon}} f(x)^*. \quad (4)$$

In fact, Figure 2 introduces a better estimate of  $\dot{\Sigma}_X$ , denoted throughout by  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ , with a smaller bias. Indeed, roughly speaking, Theorem 1 in Section 3 establishes that

$$\frac{1}{\sqrt{n}} \left\| \mathbb{E} \left[ \ddot{\Sigma}_{X,Y_{X,\epsilon}} \right] - \Sigma_\mu \right\|_F \lesssim B_{\mu,\epsilon} + \epsilon n,$$

where the expectation is over  $X, Y_{X,\epsilon}$  and  $\|\cdot\|_F$  stands for the Frobenius norm. Throughout, we will use  $\lesssim$  (and similarly  $\gtrsim, \approx$ ) to suppress universal constants and simplify the presentation. Above, the quantity  $B_{\mu,\epsilon}$  depends in a certain way on the regularity of measure  $\mu$  and function  $f(\cdot)$ , with the dependence on  $f(\cdot)$  suppressed as usual. Moreover, loosely speaking, it holds true that

$$\frac{1}{\sqrt{n}} \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \Sigma_\mu \right\|_F \lesssim B_{\mu,\epsilon} + \epsilon n + \sqrt{\frac{n}{N_{X,\epsilon}}}, \quad (5)$$

<sup>2</sup>As indicated above, we slightly abuse the standard notation by treating matrices and sets interchangeably. For example, the expression  $x \in X$  can also be interpreted as  $x$  being a column of matrix  $X$ .

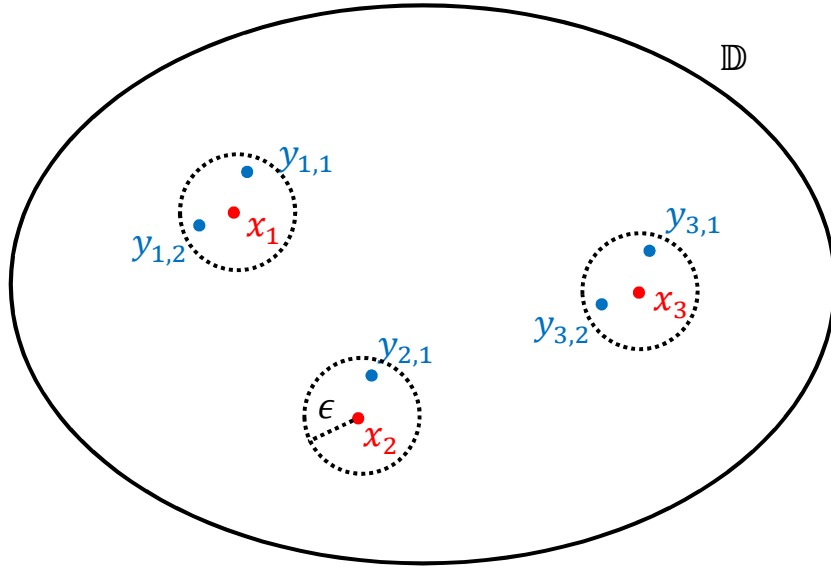


Figure 1: This diagram helps visualize the problem setup. The probability measure  $\mu$  is supported on the domain  $\mathbb{D} \subseteq \mathbb{R}^n$  of a smooth function  $f(\cdot)$ . Here,  $N = 3$  and  $X = \{x_i\}_{i=1}^N$  are drawn independently from  $\mu$ . For sufficiently small  $\epsilon$ , we let  $\mathbb{B}_{x_i, \epsilon}$  denote the  $\epsilon$ -neighborhood of each  $x_i$  and set  $\mathbb{B}_{X, \epsilon} = \cup_{i=1}^N \mathbb{B}_{x_i, \epsilon}$ . On  $\mathbb{B}_{X, \epsilon}$ ,  $\mu$  induces the conditional measure  $\mu_{X, \epsilon}$ , from which  $N_{X, \epsilon}$  points are independently drawn and collected in  $Y_{X, \epsilon} = \{y_{ij}\}_{i,j}$ . Here,  $x_1$  has  $N_{x_1, \epsilon} = 2$  neighbors in  $Y_{X, \epsilon}$  and we set  $Y_{x_1, \epsilon} = \{y_{1,j}\}_{j=1}^{N_{x_1, \epsilon}}$ . Similarly,  $Y_{x_2, \epsilon}$  and  $Y_{x_3, \epsilon}$  are formed. Note that  $Y_{X, \epsilon} = \cup_{i=1}^N Y_{x_i, \epsilon}$ . Our objective is to estimate the second-moment matrix of  $f(\cdot)$  (with respect to the probability measure  $\mu$ ) given  $\{x_i, f(x_i)\}$  and  $\{y_{ij}, f(y_{ij})\}$ .

### Estimating active subspaces:

#### Input:

- Open set  $\mathbb{D} \subseteq \mathbb{R}^n$ , equipped with probability measure  $\mu$ .
- An oracle that returns  $f(x)$  for a query point  $x \in \mathbb{D}$ .
- Neighborhood radius  $\epsilon > 0$ , sample sizes  $N$ ,  $N_{X,\epsilon}$ , and integer  $N_{X,\min,\epsilon} \leq N_{X,\epsilon}$ .

#### Output:

- $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ , as an estimate of  $\Sigma_\mu$ .

#### Body:

- Draw  $N$  random points independently from  $\mu$  and store them as the columns of  $X \in \mathbb{R}^{n \times N}$ .
- Draw  $N_{X,\epsilon}$  random points independently from  $\mu_{X,\epsilon}$  and store them as the columns of  $Y_{X,\epsilon} \in \mathbb{R}^{n \times N_{X,\epsilon}}$ . Here,  $\mu_{X,\epsilon}$  is the conditional probability measure induced by  $\mu$  on  $\mathbb{B}_{X,\epsilon} = \cup_{x \in X} \mathbb{B}_{x,\epsilon}$ . In turn,  $\mathbb{B}_{x,\epsilon} \subset \mathbb{R}^n$  is the Euclidean ball of radius  $\epsilon$  about  $x$ .
- Compute and return

$$\ddot{\Sigma}_{X,Y_{X,\epsilon}} := \frac{1}{N} \left( 1 + \frac{1 - \frac{2}{n}}{1 + \frac{2}{n}} \cdot N_{X,\min,\epsilon}^{-1} \right)^{-1} \cdot \left( \sum_{N_{x,\epsilon} \geq N_{X,\min,\epsilon}} \dot{\nabla}_{Y_{X,\epsilon}} f(x) \cdot \dot{\nabla}_{Y_{X,\epsilon}} f(x)^* - \frac{\left\| \dot{\nabla}_{Y_{X,\epsilon}} f(x) \right\|_2^2}{\left( 1 + \frac{2}{n} \right) N_{X,\min,\epsilon} + n + 1 - \frac{2}{n}} \cdot I_n \right), \quad (6)$$

where

$$\dot{\nabla}_{Y_{X,\epsilon}} f(x) := \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} \frac{f(y) - f(x)}{\|y - x\|_2} \cdot \frac{y - x}{\|y - x\|_2} \in \mathbb{R}^n. \quad (7)$$

Figure 2: The proposed algorithm for estimating the second-moment matrix of the function  $f(\cdot)$  with respect to the measure  $\mu$ .

with high probability, as described in Theorem 2 in Section 3.

Before turning to the details, consider also the following numerical example. Suppose that  $n = 100, \epsilon = 0.01$ . Let  $\mathbb{D}$  be the unit sphere in  $\mathbb{R}^n$  and take  $\mu$  to be the uniform probability measure on  $\mathbb{D}$ . Also take  $f(x) = \frac{1}{2} \sum_{i=1}^2 x[i]^2$ . For  $N$  ranging from 1 to 200 and with  $N_{X,\epsilon} \approx 2N_{X,\min,\epsilon} \cdot N = 2 \log(N) \cdot N$ , we compute  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  as given in Figure 2. The estimation error, as measured by the left-hand side of (5), is plotted in Figure 3 (after averaging over 100 trials). Note that, on average, each sample  $x \in X$  has only  $2 \log N \leq 11 \ll n = 100$  neighbors in  $Y_{X,\epsilon}$ , which makes it impossible to directly estimate  $\nabla f(x)$ . Observe also that the error decays like  $1/\sqrt{N_{X,\epsilon}}$ , as suggested by (5) and detailed later in Theorem 2 (see Section 3).

## 1.2 Contribution and Organization

The main contribution of this paper is the design and analysis of a simple algorithm to compute the second-moment matrix  $\Sigma_\mu$  of a smooth function  $f(\cdot)$  from its point samples (see (1) and Figure 2). As argued earlier and also in Section 4,  $\Sigma_\mu$  is a key quantity in ridge approximation and a number of related problems.

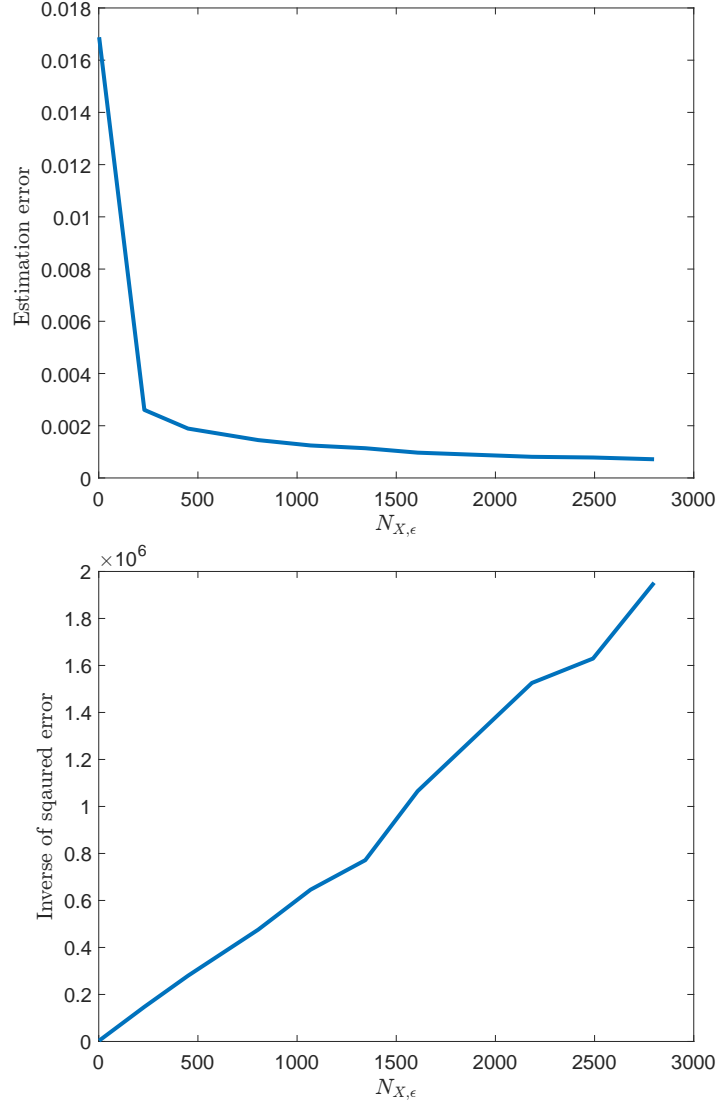


Figure 3: A numerical example for the proposed algorithm in Figure 2; see the last paragraph of Section 1.1 for details. The average estimation error, as measured by the left-hand side of (5), is plotted in the top figure versus the total number of samples used  $N_{X,\epsilon}$ . The bottom figure shows the same estimation error, squared and inverted, to emphasize the convergence rate given in (5).

The key distinction of this work is the lack of any structural assumptions (such as small rank or sparsity) on  $\Sigma_\mu$ . Imposing a specific structure on  $\Sigma_\mu$  can lead to more efficient algorithms. For example, if  $\text{rank}(\Sigma_\mu) = r$ , then one could reliably estimate  $\Sigma_\mu$  with a number of queries of the order of  $\text{poly}(r) \cdot n$  [6]; see also Section 4.

At a very high level, there is indeed a parallel between estimating the second-moment matrix of a function and estimating the covariance matrix of a random vector; our algorithm in Figure 2 might be considered as an analogue of the standard *sample covariance matrix*, adjusted to handle missing data [7]. In this context, more efficient algorithms are available for estimating, for example, the covariance matrix with a sparse inverse [8]. In this sense, we feel that this work fills an important gap in the literature of ridge approximation and perhaps dimensionality reduction by addressing the problem in full generality.

The rest of this paper is organized as follows. The problem of learning the second-moment matrix of a function is formalized in Section 2. Our approach to this problem, along with the theoretical guarantees, is described in Section 3. In Section 4, we sift through a large body of literature and summarize the relevant prior art. Proofs and technical details are deferred to Section 6 and the appendices.

## 2 Problem Statement and Approach

In this section, we formalize the problem outlined in Section 1. Consider an open set  $\mathbb{D} \subseteq \mathbb{R}^n$ , equipped with subspace Borel  $\sigma$ -algebra and probability measure  $\mu$ . We assume throughout that  $f : \mathbb{D} \rightarrow \mathbb{R}$  is twice differentiable on  $\mathbb{D}$ , and that

$$L_f := \sup_{x \in \mathbb{D}} \|\nabla f(x)\|_2 < \infty, \quad (8)$$

$$H_f := \sup_{x \in \mathbb{D}} \|\nabla^2 f(x)\|_2 < \infty, \quad (9)$$

where  $\nabla f(x) \in \mathbb{R}^n$  and  $\nabla^2 f(x) \in \mathbb{R}^{n \times n}$  are the gradient and Hessian of  $f(\cdot)$  at  $x \in \mathbb{D}$ , respectively. Moreover, for  $\epsilon > 0$ , let  $\mathbb{D}_\epsilon \subset \mathbb{D}$  denote the  $\epsilon$ -interior of  $\mathbb{D}$ , namely  $\mathbb{D}_\epsilon = \{x \in \mathbb{D} : \mathbb{B}_{x,\epsilon} \subseteq \mathbb{D}\}$ . Throughout,  $\mathbb{B}_{x,\epsilon} \subset \mathbb{R}^n$  denotes the (open) Euclidean ball of radius  $\epsilon$  centered at  $x$ .

Consider  $\Sigma_\mu \in \mathbb{R}^{n \times n}$  defined as

$$\Sigma_\mu := \mathbb{E}_x [\nabla f(x) \cdot \nabla f(x)^*] = \int_{\mathbb{D}} \nabla f(x) \cdot \nabla f(x)^* \mu(dx), \quad (10)$$

where  $\mathbb{E}_x$  computes the expectation with respect to  $x \sim \mu$ . Our objective in this work is to estimate  $\Sigma_\mu$ . To that end, consider  $N$  random points drawn independently from  $\mu$  and stored as the columns of  $X \in \mathbb{R}^{n \times N}$ . Then, as noted in Section 1.1, it is easy to verify that

$$\dot{\Sigma}_X := \frac{1}{N} \sum_{x \in X} \nabla f(x) \cdot \nabla f(x)^*, \quad (11)$$

is an unbiased estimator for  $\Sigma_\mu$  in (10). To interpret (11), recall also that we treat matrices and sets interchangeably throughout, slightly abusing the standard notation. In particular,  $x \in X$  can also be interpreted as  $x$  being a column of  $X \in \mathbb{R}^{n \times N}$ . The following standard result quantifies how well  $\dot{\Sigma}_X$  approximates  $\Sigma_\mu$ . For the sake of completeness, the proof is included in Appendix B, but also see [5, 9].

**Proposition 1.** *Let  $X \in \mathbb{R}^{n \times N}$  contain  $N$  independent samples drawn from the probability measure  $\mu$ . Then,  $\dot{\Sigma}_X$  is an unbiased estimator for  $\Sigma_\mu \in \mathbb{R}^{n \times n}$  (see (10) and (11)). Moreover, except for a probability of at most  $n^{-1}$ , it holds that*

$$\left\| \dot{\Sigma}_X - \Sigma_\mu \right\|_F \lesssim \frac{L_f^2 \log n}{\sqrt{N}}. \quad (12)$$

Since only point values of  $f(\cdot)$  are at our disposal, we cannot hope to compute  $\dot{\Sigma}_X$  directly. Instead, we will systematically generate random points near the point cloud  $X$  and then *directly* estimate  $\dot{\Sigma}_X$  by aggregating local information, as detailed next.

Given the point cloud  $X \subset \mathbb{D}$ , fix  $\epsilon > 0$ , small enough so that  $X$  is a  $2\epsilon$ -separated point cloud that belongs to the  $\epsilon$ -interior of  $\mathbb{D}$ . Formally, fix  $\epsilon \leq \epsilon_X$ , where

$$\epsilon_X := \sup \{ \epsilon' : X \subset \mathbb{D}_{\epsilon'} \text{ and } \|x - x'\|_2 \geq 2\epsilon', \forall x, x' \in X, x \neq x' \}. \quad (13)$$

Let

$$\mathbb{B}_{X,\epsilon} := \bigcup_{x \in X} \mathbb{B}_{x,\epsilon} \subseteq \mathbb{D} \quad (14)$$

denote the  $\epsilon$ -neighborhood of point cloud  $X$ . Consider the conditional probability measure on  $\mathbb{B}_{X,\epsilon}$  described as

$$\mu_{X,\epsilon} = \begin{cases} \mu / \mu(\mathbb{B}_{X,\epsilon}), & \text{inside } \mathbb{B}_{X,\epsilon}, \\ 0, & \text{outside } \mathbb{B}_{X,\epsilon}. \end{cases} \quad (15)$$

For an integer  $N_{X,\epsilon}$ , draw  $N_{X,\epsilon}$  independent random points from  $\mu_{X,\epsilon}$  and store them as the columns of  $Y_{X,\epsilon} \in \mathbb{R}^{n \times N_{X,\epsilon}}$ . Finally, an estimate of  $\dot{\Sigma}_X$  (and in turn of  $\Sigma_\mu$ ) as a function of  $X, Y_{X,\epsilon} \subset \mathbb{D}$  and  $f(\cdot)$  evaluated at these points is proposed by  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  in Figure 2.

### 3 Theoretical Guarantees

Recalling (10) and (11), how well does  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  in Figure 2 approximate  $\dot{\Sigma}_X$  and in turn  $\Sigma_\mu$ ? Parsing the answer requires introducing additional notation and imposing a certain regularity assumption on  $\mu$ . All these we set out to do now, before stating the results in Section 3.2.

For each  $x \in X$ , let the columns of  $Y_{x,\epsilon} \in \mathbb{R}^{n \times N_{x,\epsilon}}$  contain the  $\epsilon$ -neighbors of  $x$  in  $Y_{X,\epsilon}$ . In our notation, this can be written as

$$Y_{x,\epsilon} := Y_{X,\epsilon} \cap \mathbb{B}_{x,\epsilon}, \quad \#Y_{x,\epsilon} = N_{x,\epsilon}. \quad (16)$$

Because  $\epsilon \leq \epsilon_X$  is small (see (13)), these neighborhoods do not intersect, that is

$$Y_{x,\epsilon} \cap Y_{x',\epsilon} = \emptyset, \quad \forall x, x' \in X, \quad x \neq x',$$

thereby partitioning  $Y_{X,\epsilon}$  into  $\#X = N$  subsets  $\{Y_{x,\epsilon}\}_{x \in X}$ . Observe also that, conditioned on  $x \in X$  and  $N_{x,\epsilon}$ , each neighbor  $y \in Y_{x,\epsilon}$  follows the conditional probability measure described as follows:

$$y|x, N_{x,\epsilon} \sim \mu_{x,\epsilon} := \begin{cases} \mu / \mu(\mathbb{B}_{x,\epsilon}), & \text{inside } \mathbb{B}_{x,\epsilon}, \\ 0, & \text{outside } \mathbb{B}_{x,\epsilon}. \end{cases} \quad (17)$$

#### 3.1 Regularity of $\mu$

In order to introduce the regularity imposed on  $\mu$  here, consider first the special case where the domain  $\mathbb{D} \subset \mathbb{R}^n$  is bounded and  $\mu$  is the uniform probability measure on  $\mathbb{D}$ . Then, for  $\epsilon > 0$  and arbitrary  $\epsilon$ -interior point  $x \in \mathbb{D}_\epsilon$ , the conditional measure  $\mu_{x,\epsilon}$  too is the uniform measure on  $\mathbb{B}_{x,\epsilon}$  (see (17)). Draw  $y$  from  $\mu_{x,\epsilon}$ , i.e.,  $y|x \sim \mu_{x,\epsilon}$  in our notation. Then, it is easy to verify that  $y - x$  is an *isotropic* random vector, in the sense that

$$\mathbb{E}_{y|x} [(y - x)(y - x)^*] = C \cdot I_n,$$

for some factor  $C$ .<sup>3</sup> Above,  $I_n \in \mathbb{R}^{n \times n}$  is the identity matrix and  $\mathbb{E}_{y|x}[\cdot] = \mathbb{E}_y[\cdot|x]$  stands for conditional expectation, given  $x$ . A similar property plays an important role in this paper, as captured by Assumption 1 below.

**Assumption 1.** *Throughout this paper, we assume that there exist  $\epsilon_\mu, K_\mu > 0$  such that the probability measure  $\mu$  (supported on  $\mathbb{D} \subseteq \mathbb{R}^n$ ) satisfies the following requirement for an arbitrary  $\epsilon_\mu$ -interior point  $x \in \mathbb{D}_{\epsilon_\mu}$ .*

<sup>3</sup>A simple calculation shows that  $C = 1/n$ . See Appendix D.

Given  $x$ , draw  $y$  from the conditional measure on the  $\epsilon_\mu$ -neighborhood of  $x$ , namely  $y|x \sim \mu_{x,\epsilon}$  with  $\mu_{x,\epsilon}$  defined in (17). Then, for every  $\gamma_1 \geq 1$  and arbitrary (but fixed)  $v \in \mathbb{R}^n$ , it holds that

$$\Pr_{y|x} \left[ \|P_{x,y} \cdot v\|_2^2 > \gamma_1 \cdot \frac{\|v\|_2^2}{n} \right] \lesssim e^{-K_\mu \gamma_1},$$

where

$$P_{x,y} := \frac{(y-x)(y-x)^*}{\|y-x\|_2^2} \in \mathbb{R}^{n \times n} \quad (18)$$

is the orthogonal projection onto the direction of  $y-x$ . Above,  $\Pr_{y|x}[\cdot] = \Pr_y[\cdot|x]$  stands for conditional probability.

Roughly speaking, under Assumption 1,  $\mu$  is locally isotropic. Indeed, this assumption is met when  $\mu$  is the uniform probability measure on  $\mathbb{D}$ , as shown in Appendix C. Moreover, Assumption 1 is not too restrictive. One would expect that a probability measure  $\mu$ , if dominated by the uniform measure on  $\mathbb{D}$  and with a smooth Radon-Nikodym derivative, satisfies Assumption 1 when restricted to sufficiently small neighborhoods.

Assumption 1 also controls the growth of the moments of  $\|P_{x,y}v\|_2$ . Let  $\mathbb{E}^p[z] = (\mathbb{E}|z|^p)^{1/p}$  denote the  $p$ th moment of random variable  $z$ . Then, standard calculations [10] reveal that

$$\mathbb{E}_{y|x}^p \left[ \|P_{x,y} \cdot v\|_2^2 \right] \lesssim p \cdot \frac{\|v\|_2^2}{n}, \quad \forall p \geq 1. \quad (19)$$

Given the point cloud  $X$ , we also conveniently set

$$\epsilon_{\mu,X} := \min[\epsilon_\mu, \epsilon_X]. \quad (\text{see Assumption 1 and (13)}) \quad (20)$$

We are now in position to present the the main results.

### 3.2 Theoretical Guarantees

With the setup detailed in Section 2, we now summarize our findings here. In Theorems 1 and 2 below, for a fixed point cloud  $X$ , we focus on how well  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  in Figure 2 approximates  $\Sigma_\mu$ . Then, in the ensuing remarks, we combine these results with Proposition 1 to see how well  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  approximates  $\Sigma_\mu$ . We now turn to the details.

Theorem 1 below states that  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  can be a nearly unbiased estimator of  $\dot{\Sigma}_X$  given  $X$  (see (11)). The proof is given in Section 6.1. Throughout,  $\mathbb{E}_{z_1|z_2}[\cdot] = \mathbb{E}_{z_1}[\cdot|z_2]$  stands for conditional expectation over  $z_1$  and conditioned on  $z_2$  for random variables  $z_1, z_2$ .

**Theorem 1. (Bias)** Consider an open set  $\mathbb{D} \subseteq \mathbb{R}^n$  equipped with probability measure  $\mu$ , and consider a twice differentiable function  $f : \mathbb{D} \rightarrow \mathbb{R}$  satisfying (8,9). Assume that the (fixed) columns of  $X \in \mathbb{R}^{n \times N}$  belong to  $\mathbb{D}$ , namely  $X \subset \mathbb{D}$  in our notation. Fix also  $\epsilon \in (0, \epsilon_{\mu,X}]$  (see (20)). For integers  $N, N_{X,\epsilon}$  and  $N_{X,\min,\epsilon} \leq N_{X,\epsilon}$ , assume also that

$$N_{X,\epsilon} \geq \frac{N_{X,\min,\epsilon} N}{\rho_{\mu,X,\epsilon}}, \quad N_{X,\min,\epsilon} \gtrsim \log^2 N, \quad \log(1/\epsilon) \gtrsim \frac{\log(K_\mu n)}{\log N_{X,\epsilon}}, \quad (21)$$

where

$$\rho_{\mu,X,\epsilon} := N \cdot \min_{x \in X} \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})}. \quad (22)$$

Then, the output of the algorithm in Figure 2, namely  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ , satisfies

$$\frac{1}{\sqrt{n}} \left\| \mathbb{E}_{Y_{X,\epsilon}|X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \dot{\Sigma}_X \right\|_F \lesssim B_{\mu,\epsilon} + n\epsilon \sqrt{\log(1/\epsilon)} (L_f + H_f)^2 \sqrt{\frac{\log N_{X,\epsilon}}{K_\mu}} + \frac{L_f^2 n^{\frac{3}{2}}}{N^{10}}, \quad (23)$$

where  $B_{\mu,\epsilon}$  is given explicitly in (46).



A few remarks are in order.

**Remark 1. (Discussion)** Theorem 1 describes how well  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  approximates  $\dot{\Sigma}_X$ , in expectation. To form a better understanding of this result, let us first study the conditions listed in (21).

- The requirements on  $N_{X,\epsilon}$  and  $N_{X,\min,\epsilon}$  in (21) ensure that every  $x \in X$  has sufficiently many neighbors in  $Y_{X,\epsilon}$ , i.e.,  $N_{x,\epsilon}$  is large enough for all  $x$ . For example, if  $\mathbb{D} \subset \mathbb{R}^n$  is bounded and  $\mu$  is the uniform probability measure on  $\mathbb{D}$ , then  $\rho_{\mu,X,\epsilon} = 1$ . We might then take  $N_{X,\min,\epsilon} \approx \log^2 N$  and  $N_{X,\epsilon} \gtrsim N \log^2 N$  so that (23) holds with the total of  $N_{X,\epsilon} = O(N \log^2 N)$  samples.
- The upper bound on the neighborhood radius  $\epsilon$  in the last inequality of (21) is not too restrictive. For example, suppose that  $K_\mu = 1$  and  $N_{X,\epsilon} = O(N)$ . Then, roughly speaking, the third inequality in (21) requires that  $\epsilon \lesssim n^{-\frac{C}{\log N}}$  for some factor  $C$ . Indeed, as we will see shortly, the bound on the estimation error in (23) is often only useful for values of  $\epsilon$  well within this range.

Let us next interpret the bound on the bias in (23).

- The term  $B_{\mu,\epsilon}$  is given explicitly in (46); it depends on the probability measure  $\mu$  and the function  $f(\cdot)$ . In fact, the bound in (23) might be sharpened by replacing  $B_{\mu,\epsilon}$  with a somewhat more involved expression, but we opted for the present form for simplicity. As a special case, however, suppose that  $\mu$  is the uniform probability measure on a bounded and open set  $\mathbb{D}$ , and that every  $x \in X$  has the same number of neighbors  $N_{x,\epsilon} = N_{X,\epsilon}/N$  within  $Y_{X,\epsilon}$ . Then it is evident from the proof of Theorem 1 that (23) can in fact be sharpened by setting  $B_{\mu,\epsilon} = 0$ . Indeed,  $B_{\mu,\epsilon}$  captures the non-uniformity of  $\mu$ ; the more isotropic  $\mu$  is (in the sense described in Assumption 1), the smaller  $B_{\mu,\epsilon}$  would be.
- The second term on the right-hand side of (23) is due to the higher order terms in the Taylor expansion of  $f(\cdot)$ . In particular, when  $f(\cdot)$  is sufficiently smooth,  $L_f + H_f$  is small and the second term is negligible. This term also suggests taking  $\epsilon$  proportional to  $(L_f + H_f)^{-2} n^{-\frac{3}{2}}$ .
- The last term on the right-hand side of (23) accounts for the possibility of  $Y_{X,\epsilon}$  not being distributed evenly. That is, it accounts for the possibility of some samples in  $X$  having too few neighbors in  $Y_{X,\epsilon}$ . We point out that there is nothing specific about the 10th power of  $N$  in the denominator and a more general expression is given in the proof.

**Remark 2. (Sampling strategy)** In Figure 2,  $N_{X,\epsilon}$  points are independently drawn from the conditional probability measure on the  $\epsilon$ -neighborhood of the point cloud  $X$  and then stored as the columns of  $Y_{X,\epsilon}$ , namely

$$Y_{X,\epsilon} \stackrel{\text{i.i.d.}}{\sim} \mu_{X,\epsilon}. \quad (\text{see (15)}) \quad (24)$$

This sampling strategy appears to best fit our fixed budget of  $N_{X,\epsilon}$  samples, as it “prioritizes” the areas of  $\mathbb{D}$  with larger “mass.” For example, suppose that  $\mu(dx) \gg \mu(dx')$  and  $\nabla f(x) \approx \nabla f(x')$  for a pair  $x, x' \in \mathbb{D}$ . Then,  $\nabla f(x) \nabla f(x)^* \mu(dx) \gg \nabla f(x') \nabla f(x')^* \mu(dx')$ , suggesting that a larger weight should be placed on  $x$  rather than  $x'$  when estimating  $\Sigma_\mu$  (see (10)). In the same scenario, assume naturally that  $\mu(\mathbb{B}_{x,\epsilon}) \gg \mu(\mathbb{B}_{x',\epsilon})$ , so that it is more likely to sample from the neighborhood of  $x$  than  $x'$ . Then, given a fixed budget of  $N_{X,\epsilon}$  samples, it is highly likely that  $N_{x,\epsilon} \gg N_{x',\epsilon}$ . That is,  $x$  likely has far more  $\epsilon$ -neighbors in  $Y_{X,\epsilon}$  compared to  $x'$ . Loosely speaking then, the contribution of  $x$  to  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  is calculated more accurately than that of  $x'$ . In other words, the sampling strategy used in Figure 2 indeed assigns more weight to areas of  $\mathbb{D}$  with larger mass.

In contrast, to generate  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ , assume that  $Y_{X,\epsilon}$  were instead drawn from the uniform measure on the  $\epsilon$ -neighborhood of  $X$ , i.e.,

$$Y_{X,\epsilon} \stackrel{\text{i.i.d.}}{\sim} \text{Uniform}(\mathbb{B}_{X,\epsilon}). \quad (\text{see (14)}) \quad (25)$$

Revisiting the proof of Theorem 1 (in particular, proof of Lemma 3) reveals that making the above modification in Figure 2 replaces  $B_{\mu,\epsilon}$  in (23) with a smaller term proportional to  $L_f^2 / N_{X,\min,\epsilon}$ , because  $B_{\mu,\epsilon}$  effectively measures how locally isotropic  $\mu$  is. However, the modification in (25) will result in *not* sampling the neighborhood  $B_{\mu,\epsilon}$  according to the weight (or importance) assigned by  $\mu$ . As a result, nearly equal resources will be spent on estimating  $\nabla f(x) \nabla f(x)^*$  both where  $\|\nabla f(x)\|_2 \gg 1$  and where  $\|\nabla f(x)\|_2 \ll 1$ . It must however be noted that we did not attempt to quantify the qualitative comparisons between sampling strategies in this remark.

**Remark 3. (Proof strategy)** At a high level, the analysis handles the possible non-uniformity of the measure  $\mu$  and higher order terms in  $f(\cdot)$  by introducing quantities that are simpler to work with but are similar to  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ . Moreover, if  $N_{X,\epsilon}$  is sufficiently large, then each  $x \in X$  has many neighbors in  $Y_{X,\epsilon}$  and this observation aids the analysis. The rest of the calculations, in effect, remove the estimation bias of  $\dot{\Sigma}_{X,Y_{X,\epsilon}}$  (see (4)) in order to arrive at  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ .

Our second result, proved in Section 6.2, is a finite-sample bound for  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ . Theorem 2 states that  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  can reliably estimate  $\dot{\Sigma}_X$ , with high probability.

**Theorem 2. (Finite-sample bound)** *Under the same setup as in Theorem 1 and for  $\gamma_2 \geq 1$ , it holds that*

$$\frac{1}{\sqrt{n}} \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \dot{\Sigma}_X \right\|_F \lesssim B_{\mu,\epsilon} + \epsilon n L_f H_f \sqrt{\gamma_2 \log N_{X,\epsilon}} + \gamma_2^2 K_\mu L_f^2 \log^3(n N_{X,\epsilon}) \sqrt{\frac{n}{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}}, \quad (26)$$

except with a probability of

$$O(n^{2-\log \gamma_2} + N^{1-2 \log K_\mu - \log \gamma_2}), \quad (27)$$

and provided that

$$N_{X,\epsilon} \geq \frac{N_{X,\min,\epsilon} N}{\rho_{\mu,X,\epsilon}}, \quad N_{X,\min,\epsilon} \gtrsim \gamma_2^2 K_\mu^2 \log^4(n N_{X,\epsilon}). \quad (28)$$

Here,  $O(\cdot)$  is the standard Big-O notation.

**Remark 4. (Discussion)** Let us dissect the estimation error, namely the right-hand side of (26). As discussed in Remark 1,  $B_{\mu,\epsilon}$  in effect captures the non-uniformity of measure  $\mu$ . In particular, the right-hand side of (26) can be sharpened by setting  $B_{\mu,\epsilon} = 0$  in the setting described in that remark.

Similar to Remark 1, the second term on the right-hand side of (26) reflects the higher order terms in the Taylor expansion of  $f(\cdot)$  and, roughly speaking, suggests taking  $\epsilon$  proportional to  $n^{-\frac{3}{2}}$ . The last term on the right-hand side of (26) sets the convergence rate. In particular, if  $\mu$  is the uniform probability measure on a bounded and open set  $\mathbb{D}$ , then  $\rho_{\mu,X,\epsilon} = 1$ . Then, ignoring the logarithmic factors, the convergence rate is approximately  $1/\sqrt{N_{X,\epsilon}}$ . See Section 4 for comparison with the existing literature.

Next, let us study the requirements in (28). Assuming uniform measure and that  $\rho_{\mu,X,\epsilon} = \gamma_2 = K_\mu = O(1)$  in order to gain insight, (28) loosely requires that  $N_{X,\epsilon} \gtrsim N \log^4(nN)$ . That is, to successfully estimate  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ , every sample  $x \in X$  should have  $O(\log^4(nN))$  neighbors in  $Y_{X,\epsilon}$ . We suspect that this logarithmic dependence can be improved.

**Remark 5. (Estimating  $\Sigma_\mu$ )** We remark that combining Theorem 2 with Proposition 1 yields

$$\frac{\|\ddot{\Sigma}_{X,Y_{X,\epsilon}} - \Sigma_\mu\|_F}{\sqrt{n}} \lesssim B_{\mu,\epsilon} + \epsilon n L_f H_f \sqrt{\gamma_2 \log N_{X,\epsilon}} + \gamma_2^2 K_\mu L_f^2 \log^3(n N_{X,\epsilon}) \sqrt{\frac{n}{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} + \frac{L_f^2 \log n}{\sqrt{nN}} =: \Delta_{\mu,\epsilon}, \quad (29)$$

with high probability, therefore quantifying how well the algorithm in Figure 2 estimates the second-moment matrix of  $f(\cdot)$ .

**Remark 6. (Proof strategy)** The estimation error here is decomposed into “diagonal” and “off-diagonal” terms. The diagonal term, we find, can be written as a sum of independent random matrices and controlled by applying a standard Bernstein inequality. The off-diagonal term, however, is a second-order chaos and requires additional care.

**Remark 7. (Improvements)** In combination with Weyl’s inequality, (29) might be used to control the distance between the spectrum of  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  and that of  $\Sigma_\mu$ . Likewise, given an integer  $r \leq n$ , standard perturbation results [11] might be deployed to measure the principal angle between the span of the leading  $r$  eigenvectors of  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  and an  $r$ -dimensional active subspace of  $f(\cdot)$ . To obtain the sharpest bounds, both these improvements would require controlling the spectral norm of  $\ddot{\Sigma}_{X,Y_{X,\epsilon}} - \Sigma_\mu$  rather than its Frobenius norm (which is bounded in Theorem 2 above). Controlling the spectral norm of the error appears to be more difficult to achieve. Indeed, the Frobenius norm of the estimation error can be decomposed into sum of (squared) scalar chaos random variables, for which a comprehensive theory is known, whereas it is not clear how to approach the spectral norm directly. As an aside, let us point out that the spectrum of  $\Sigma_\mu$  in relation to  $f(\cdot)$  has been studied in [3, 6].

## 4 Related Work

As argued in Section 1, the second-moment matrix (or its leading eigenvectors) is of particular importance in the context of ridge approximation, for which a large body of work exists in the literature of approximation theory [12, 13, 2, 14, 15, 16, 17, 18, 19, 20]. More specifically, in [3] for instance, the authors develop an algorithm to learn an active subspace of a function (namely, the span of the leading eigenvectors of the second-moment matrix) when its basis vectors are (nearly) sparse. The sparsity assumption was later removed in [6, 21] and replaced with the notion of a low-dimensional active subspace. We emphasize that these models allow for algorithms with better sample complexities compared to Theorem 2 which, in contrast, holds without any assumption on the second-moment matrix. In this sense, the present work fills a gap in the literature of ridge approximation; see also Section 1.2.

Another closely related topic in high-dimensional statistics and machine learning is *projection pursuit* where, spurred by the interest in *generalized additive models* [22], the aim is to approximate  $f(\cdot)$  with functions of the form  $\sum_i h_i(u_i^* x)$  [23, 24, 25]. Further connections with neural networks are studied in [26, 27]. See also [28, 29] for relation with the Gaussian process regression and uncertainty quantification. *Sufficient dimension reduction* and related topics [30, 31, 32, 33, 22, 34, 35, 36] is yet another related line of work in statistics. In this context, the typical assumption is that  $f(x) = h(U^* x)$  is an exact ridge function, observed under additive noise. The objective is then to estimate  $U = \text{span}(U)$ , known as the *effective subspace for regression* in this literature.

Finding the second-moment matrix of a function is also closely related to covariance estimation (see (1)), which is widely studied in modern statistics often under various structural assumptions on the covariance matrix, e.g., sparsity of its inverse [37, 38, 39, 40, 41]. In this context, it appears that [42, 43, 44, 45] are the most relevant to the present work, in part because of their lack of any structural assumptions. For the sake of brevity, we focus on [42], which offers an unbiased estimator for the covariance matrix of a random vector  $x$  given few measurements of multiple realizations of  $x$  in the form of  $\{\Phi_i x_i\}_i$  for low-dimensional (and uniformly random) orthogonal projection matrices  $\{\Phi_i\}_i$ . The fact that the distributions involving  $\Phi_i x_i$  admit closed forms enables the authors to provide a comprehensive theory for their estimator without explicitly introducing chaos random variables into their analysis. It is important to point out that, by design, the estimator in [42] is not applicable to our setup.<sup>4</sup> Our framework might be interpreted as sum of rank-1 projections, which introduces chaos random variables into our estimator (and substantially complicates the analysis). To further complicate the matters, the probability measure  $\mu$  on  $\mathbb{D}$  is not necessarily uniform; we cannot hope to explicitly determine the distribution of the crucial components of the estimator. Instead, we rely on the standard tools in empirical processes to control the chaos terms. We note that the convergence rate in Theorem 2 here appears to match the rates obtained in [42, Theorem 3]. It is also worth including a few other key pointers here [7, 46, 47].

In computational sciences, the input/output relation in a complex process is often modelled with a cheap surrogate for the sake of making fast and affordable predictions. In this context, the applications of ridge approximation are well-documented and we refer to [48, 49, 50, 51, 52, 9, 53, 54].

Yet another related field is matrix completion and recovery [55, 56, 57] and subspace estimation from data with erasure [58], where typically a low-rank structure is imposed. Lastly, in numerical linear algebra, random projections are increasingly used to facilitate matrix operations [59, 60, 61]. As a result, a very similar mathematical toolbox is used in this line of research.

## 5 Acknowledgements

AE would like to thank Hemant Tyagi for many interesting conversations regarding ridge approximation. Some parts of this project were supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1. MBW was partially supported by NSF grant CCF-1409258 and NSF CAREER grant CCF-1149225. PGC was partially supported by the U.S. Department of Energy Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under award DE-SC0011077 and the Defense Advanced Research Projects Agency's Enabling Quantification of Uncertainty in Physical Systems. RAW was supported by NSF CAREER grant CCF-1255631.

<sup>4</sup>Forward Euler method will effectively replace  $\Phi_t x_t$  in  $\hat{\Sigma}_1$  in [42, Section 3] with a sum of rank-1 projections of  $x_t$ .

## 6 Theory

This section contains the proofs of the two main results of this paper.

### 6.1 Proof of Theorem 1

Let us begin by outlining the proof strategy.

- First, we introduce a new quantity:  $\ddot{\Sigma}_{X,Y_{X,\epsilon}} \in \mathbb{R}^{n \times n}$ . Conditioned on a certain “good” event  $\mathcal{E}_1$ ,  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  is easier to work with than  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ .
- Then, for fixed  $X \subset \mathbb{D}_\epsilon$ , we define another “good” event  $\mathcal{E}_2$  where each  $x \in X$  has many neighbors in  $Y_{X,\epsilon}$ . Lemma 2 below shows that  $\mathcal{E}_2$  is very likely to happen if  $N_{X,\epsilon} = \#Y_{X,\epsilon}$  is large enough. Conditioned on the event  $\mathcal{E}_2$ , Lemma 3 below shows that  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  is a nearly unbiased estimator of  $\dot{\Sigma}_X$ :

$$\mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] \approx \dot{\Sigma}_X. \quad (30)$$

- Lastly, we remove the conditioning on  $\mathcal{E}_1 \cap \mathcal{E}_2$  to complete the proof of Theorem 1.

We now turn to the details and introduce  $\ddot{\Sigma}_{X,Y_{X,\epsilon}} \in \mathbb{R}^{n \times n}$ :

$$\ddot{\Sigma}_{X,Y_{X,\epsilon}} := \frac{1}{N} \left( 1 + \frac{1 - \frac{2}{n}}{1 + \frac{2}{n}} \cdot N_{X,\min,\epsilon}^{-1} \right)^{-1} \cdot \left( \sum_{N_{x,\epsilon} \geq N_{X,\min,\epsilon}} \ddot{\nabla}_{Y_{X,\epsilon}} f(x) \cdot \ddot{\nabla}_{Y_{X,\epsilon}} f(x)^* - \frac{\sum_{N_{x,\epsilon} \geq N_{X,\min,\epsilon}} \|\ddot{\nabla}_{Y_{X,\epsilon}} f(x)\|_2^2}{(1 + \frac{2}{n}) N_{X,\min,\epsilon} + n + 1 - \frac{2}{n}} \cdot I_n \right), \quad (31)$$

Here,

$$\ddot{\nabla}_{Y_{X,\epsilon}} f(x) := \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{X,\epsilon}} P_{x,y} \cdot \nabla f(x) \in \mathbb{R}^n, \quad (32)$$

and  $P_{x,y} \in \mathbb{R}^{n \times n}$  is the orthogonal projection onto the direction of  $y - x$ . In order to relate  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  to  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$ , we invoke the following result, proved in Appendix E.

**Lemma 1.** Fix  $X$  and  $\epsilon \in (0, \epsilon_{\mu,X}]$ . It holds that

$$\frac{1}{\sqrt{n}} \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \ddot{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F \lesssim \epsilon L_f H_f n^{\frac{3}{2}}. \quad (33)$$

Moreover, consider the event

$$\mathcal{E}_1 := \left\{ Y_{X,\epsilon} \mid \max_{x \in X} \max_{y \in Y_{X,\epsilon}} \|P_{x,y} \cdot \nabla f(x)\|_2^2 \leq \frac{Q_{X,\epsilon} L_f^2}{n} \right\}, \quad (34)$$

for  $Q_{X,\epsilon} > 0$  to be set later. Then, conditioned on the event  $\mathcal{E}_1$ , it holds that

$$\frac{1}{\sqrt{n}} \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \ddot{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F \lesssim \epsilon L_f H_f Q_{X,\epsilon}^{\frac{1}{2}} n. \quad (35)$$

Thanks to Assumption 1, the event  $\mathcal{E}_1$  is very likely to happen for the right choice of  $Q_{X,\epsilon}$ . Indeed, if we set  $Q_{X,\epsilon} = \gamma_2 \log N_{X,\epsilon}$  for  $\gamma_2 \geq 1$ , then

$$\Pr_{Y_{X,\epsilon}|X} [\mathcal{E}_1^C] \leq N_{X,\epsilon}^{1-K_\mu \gamma_2}, \quad (36)$$

which follows from (8) and an application of the union bound (similar to the slightly more general result in Lemma 8).

Roughly speaking, In light of Lemma 1,  $\ddot{\Sigma}_{X,Y_{X,\epsilon}} \approx \ddot{\Sigma}_{X,Y_{X,\epsilon}}$ . It therefore suffices to study the bias of  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  in the sequel. As suggested earlier, if  $\#Y_{X,\epsilon} = N_{X,\epsilon}$  is sufficiently large, then every  $x \in X$  will likely have many neighbors in  $Y_{X,\epsilon}$ , namely  $\#Y_{x,\epsilon} = N_{x,\epsilon} \gg 1$  for every  $x \in X$ . This claim is formalized below and proved in Appendix F.

**Lemma 2.** Fix  $X$  and  $\epsilon \in (0, \epsilon_X]$ . With  $\gamma_3 \geq 1$ , assume that

$$N_{X,\epsilon} \gtrsim \frac{\gamma_3^2 \log^2 N \cdot \mu(\mathbb{B}_{X,\epsilon})}{\min_{x \in X} \mu(\mathbb{B}_{x,\epsilon})}. \quad (37)$$

Then, except with a probability of at most  $N^{1-\log \gamma_3}$ , it holds that

$$\frac{1}{2} \cdot \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} N_{X,\epsilon} \leq N_{x,\epsilon} \leq \frac{3}{2} \cdot \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} N_{X,\epsilon}, \quad \forall x \in X. \quad (38)$$

To use Lemma 2 here, we proceed as follows. For  $\gamma_3 \geq 1$ , suppose that

$$N_{X,\min,\epsilon} \gtrsim \gamma_3^2 \log^2 N, \quad (39)$$

and consider the event

$$\mathcal{E}_2 := \bigcap_{x \in X} \left\{ Y_{x,\epsilon} \mid N_{x,\epsilon} \geq \frac{1}{2} \cdot \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} N_{X,\epsilon} \geq N_{X,\min,\epsilon} \right\}, \quad (40)$$

where, in particular, each  $x \in X$  has at least  $N_{X,\min,\epsilon}$  neighbors in  $Y_{x,\epsilon}$ . In light of Lemma 2,  $\mathcal{E}_2$  is very likely to happen. To be specific,

$$\Pr_{Y_{X,\epsilon}|X} [\mathcal{E}_2^C] \leq N^{1-\log \gamma_3}, \quad (41)$$

provided that

$$N_{X,\epsilon} \gtrsim \frac{N_{X,\min,\epsilon} \cdot \mu(\mathbb{B}_{X,\epsilon})}{\min_{x \in X} \mu(\mathbb{B}_{x,\epsilon})} = \frac{N_{X,\min,\epsilon} N}{\rho_{\mu,X,\epsilon}}, \quad (\text{see (39)}) \quad (42)$$

where we conveniently defined

$$\rho_{\mu,X,\epsilon} = N \cdot \min_{x \in X} \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})}. \quad (43)$$

Conditioned on the event  $\mathcal{E}_2$ ,  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  in (31) takes the following simplified form:

$$\ddot{\Sigma}_{X,Y_{X,\epsilon}} = \frac{1}{N} \left( 1 + \frac{1 - \frac{2}{n}}{1 + \frac{2}{n}} \cdot N_{X,\min,\epsilon}^{-1} \right)^{-1} \left( \sum_{x \in X} \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \ddot{\nabla}_{Y_{x,\epsilon}} f(x)^* - \frac{\sum_{x \in X} \left\| \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2^2}{\left( 1 + \frac{2}{n} \right) N_{X,\min,\epsilon} + n + 1 - \frac{2}{n}} \cdot I_n \right). \quad (44)$$

Using the above simplified form, we will prove the following result in Appendix G. Roughly speaking it states that, conditioned on the event  $\mathcal{E}_2$ ,  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  is a nearly-unbiased estimator of  $\dot{\Sigma}_X$ .

**Lemma 3.** Fix  $X$  and  $\epsilon \in (0, \epsilon_{\mu,X}]$ . Then, it holds that

$$\frac{1}{\sqrt{n}} \left\| \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \dot{\Sigma}_X \right\|_F \lesssim B_{\mu,\epsilon}, \quad (45)$$

where

$$B_{\mu,\epsilon} := \frac{(B'_{\mu,\epsilon} + 1)^2 L_f^2}{\sqrt{n}} + \frac{B''_{\mu,\epsilon} + L_f^2}{N_{X,\min,\epsilon}}, \quad (46)$$

$$B'_{\mu,\epsilon} := n \cdot \sup_{x \in \mathbb{D}_\epsilon} \left\| \mathbb{E}_{y|x} [P_{x,y}] - \frac{I_n}{n} \right\|, \quad (y|x \sim \mu_{x,\epsilon})$$

$$B''_{\mu,\epsilon} := n^{\frac{3}{2}} \cdot \sup_{x \in \mathbb{D}_\epsilon} \left\| \mathbb{E}_{y|x} [P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}] - \left( \frac{2 \nabla f(x) \nabla f(x)^*}{n(n+2)} + \frac{\|\nabla f(x)\|_2^2}{n(n+2)} \cdot I_n \right) \right\|_F, \quad (y|x \sim \mu_{x,\epsilon}).$$

In particular, suppose that  $\mu$  is the uniform probability measure on  $\mathbb{D}$ , and that  $N_{x,\epsilon} = N_{x',\epsilon}$  for every pair  $x, x' \in X$ . Then, conditioned on  $\mathcal{E}_2$ ,  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  is an unbiased estimator of  $\dot{\Sigma}_X$ .

Next, we remove the conditioning on the event  $\mathcal{E}_2$ , with the aid of the following bounds:

$$\begin{aligned}
\left\| \dot{\Sigma}_X \right\|_F &\lesssim L_f^2, \quad (\text{see (11) and (8)}) \\
\left\| \ddot{\nabla}_{Y_{X,\epsilon}} f(x) \right\|_2 &\leq n L_f, \quad \forall x \in X, \quad (\text{see (32) and (8)}) \\
\left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F &\lesssim n^2 L_f^2. \quad (\text{see (31) and (8)})
\end{aligned} \tag{47}$$

Then, we write that

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \left\| \mathbb{E}_{Y_{X,\epsilon}|X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \dot{\Sigma}_X \right\|_F \\
&= \frac{1}{\sqrt{n}} \left\| \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] \cdot \Pr_{Y_{X,\epsilon}|X} [\mathcal{E}_2] + \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2^C,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] \cdot \Pr_{Y_{X,\epsilon}|X} [\mathcal{E}_2^C] - \dot{\Sigma}_X \right\|_F \\
&\leq \frac{1}{\sqrt{n}} \left\| \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \dot{\Sigma}_X \right\|_F \cdot \Pr_{Y_{X,\epsilon}|X} [\mathcal{E}_2] + \frac{1}{\sqrt{n}} \left\| \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2^C,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \dot{\Sigma}_X \right\|_F \cdot \Pr_{Y_{X,\epsilon}|X} [\mathcal{E}_2^C] \\
&\leq \frac{1}{\sqrt{n}} \left\| \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \dot{\Sigma}_X \right\|_F \cdot \Pr_{Y_{X,\epsilon}|X} [\mathcal{E}_2] + \frac{1}{\sqrt{n}} \left( \sup \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F + \sup \left\| \dot{\Sigma}_X \right\|_F \right) \cdot \Pr_{Y_{X,\epsilon}|X} [\mathcal{E}_2^C] \\
&\leq \frac{1}{\sqrt{n}} \left\| \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \dot{\Sigma}_X \right\|_F + n^{\frac{3}{2}} L_f^2 \cdot N^{1-\log \gamma_3} \quad (\text{see (47) and (41)}) \\
&\lesssim B_{\mu,\epsilon} + n^{\frac{3}{2}} L_f^2 \cdot N^{1-\log \gamma_3}, \quad (\text{see Lemma 3 and (11)})
\end{aligned} \tag{48}$$

which, to reiterate, holds with  $N_{X,\min,\epsilon} \gtrsim \gamma_3^2 \log^2 N$  and under (42). Lastly, we reintroduce  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  by invoking Lemma 1 as follows:

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \left\| \mathbb{E}_{Y_{X,\epsilon}|X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}} - \ddot{\Sigma}_{X,Y_{X,\epsilon}}] \right\|_F \\
&\leq \frac{1}{\sqrt{n}} \left\| \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}} - \ddot{\Sigma}_{X,Y_{X,\epsilon}}] \right\|_F \cdot \Pr_{Y_{X,\epsilon}|X} [\mathcal{E}_1] \\
&\quad + \frac{1}{\sqrt{n}} \left( \sup \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F + \sup \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F \right) \cdot \Pr_{Y_{X,\epsilon}|X} [\mathcal{E}_1^C] \quad (\text{similar to (48)}) \\
&\lesssim \epsilon L_f H_f Q_{X,\epsilon}^{\frac{1}{2}} n + \frac{1}{\sqrt{n}} \left( \sup \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F + \sup \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F \right) \cdot \Pr_{Y_{X,\epsilon}|X} [\mathcal{E}_1^C] \quad (\text{see (35)}) \\
&\lesssim \epsilon L_f H_f (\gamma_2 \log N_{X,\epsilon})^{\frac{1}{2}} n \\
&\quad + \frac{1}{\sqrt{n}} \left( \sup \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F + \sup \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F \right) \cdot N_{X,\epsilon}^{1-K_\mu \gamma_2} \quad (\text{with the choice of } Q_{X,\epsilon} \text{ in (36)}) \\
&\leq \epsilon L_f H_f (\gamma_2 \log N_{X,\epsilon})^{\frac{1}{2}} n \\
&\quad + \left( \sup \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \ddot{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F + 2 \sup \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F \right) \cdot N_{X,\epsilon}^{1-K_\mu \gamma_2} \quad (\text{triangle inequality}) \\
&\leq \epsilon L_f H_f (\gamma_2 \log N_{X,\epsilon})^{\frac{1}{2}} n \\
&\quad + \frac{1}{\sqrt{n}} (\epsilon L_f H_f n^2 + 2 L_f^2 n^2) \cdot N_{X,\epsilon}^{1-K_\mu \gamma_2} \quad (\text{see (33) and (47)}) \\
&\lesssim \epsilon L_f H_f (\gamma_2 \log N_{X,\epsilon})^{\frac{1}{2}} n + L_f (L_f + H_f) n^{\frac{3}{2}} \cdot N_{X,\epsilon}^{1-K_\mu \gamma_2} \quad (\text{when } \epsilon \lesssim 1) \\
&\lesssim n \epsilon \sqrt{\log(1/\epsilon)} L_f (L_f + H_f) \sqrt{\frac{\log N_{X,\epsilon}}{K_\mu}}. \quad \left( \text{if } \gamma_2 \approx \frac{\log(1/\epsilon)}{K_\mu} \text{ and } \epsilon \lesssim e^{-\frac{C_1 \log(K_\mu n)}{\log N_{X,\epsilon}}} \right)
\end{aligned} \tag{49}$$

Combining the above bound with (48) yields that

$$\begin{aligned}
& \frac{1}{\sqrt{n}} \left\| \mathbb{E}_{Y_{X,\epsilon}|X} \left[ \ddot{\Sigma}_{X,Y_{X,\epsilon}} \right] - \dot{\Sigma}_X \right\|_F \\
& \leq \frac{1}{\sqrt{n}} \left\| \mathbb{E}_{Y_{X,\epsilon}|X} \left[ \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \ddot{\Sigma}_{X,Y_{X,\epsilon}} \right] \right\|_F + \frac{1}{\sqrt{n}} \left\| \mathbb{E}_{Y_{X,\epsilon}|X} \left[ \ddot{\Sigma}_{X,Y_{X,\epsilon}} \right] - \dot{\Sigma}_X \right\|_F \quad (\text{triangle inequality}) \\
& \lesssim n\epsilon \sqrt{\log(1/\epsilon)} L_f (L_f + H_f) \sqrt{\frac{\log N_{X,\epsilon}}{K_\mu}} + B_{\mu,\epsilon} + n^{\frac{3}{2}} L_f^2 N^{1-\log \gamma_3}. \quad (\text{see (48) and (49)})
\end{aligned} \tag{50}$$

This completes the proof of Theorem 1 after taking  $\gamma_3 \approx 1$ .

## 6.2 Proof of Theorem 2

At a high level, the proof strategy here matches that of Theorem 1. First, we replace  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  with the simpler quantity  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  defined in (31). More specifically, in light of Lemma 1, it suffices to study  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  in the sequel.

Next, for  $N_{X,\min,\epsilon} > 0$  to be set later, recall the “good” event  $\mathcal{E}_2$  in (40) whereby every  $x \in X$  has at least  $N_{X,\min,\epsilon}$  neighbors in  $Y_{X,\epsilon}$ . Conditioned on the event  $\mathcal{E}_2$ ,  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  takes the simpler form of (44), using which we prove the following result in Appendix H.

**Lemma 4.** Fix  $X$  and  $\epsilon \in (0, \epsilon_{\mu,X}]$ . Conditioned on the event  $\mathcal{E}_2$ , it holds that

$$\frac{1}{\sqrt{n}} \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \dot{\Sigma}_X \right\|_F \lesssim B_{\mu,\epsilon} + \gamma_7 \gamma_2 K_\mu L_f^2 \log^3(nN_{X,\epsilon}) \sqrt{\frac{n}{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}}, \tag{51}$$

except with a probability of at most

$$(nN_{X,\epsilon})^{(1-K_\mu^2 \gamma_2) \log(nN_{X,\epsilon})} + n^{2-\log \gamma_7},$$

and provided that

$$N_{X,\min,\epsilon} \gtrsim \gamma_2^2 K_\mu^2 \log^4(nN_{X,\epsilon}). \tag{52}$$

We next remove the conditioning on the event  $\mathcal{E}_2$  by writing that

$$\begin{aligned}
& \Pr_{Y_{X,\epsilon}|X} \left[ \frac{1}{\sqrt{n}} \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \dot{\Sigma}_X \right\|_F \gtrsim B_{\mu,\epsilon} + \gamma_7 \gamma_2 K_\mu L_f^2 \log^3(nN_{X,\epsilon}) \sqrt{\frac{n}{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \right] \\
& \leq \Pr_{Y_{X,\epsilon}|\mathcal{E}_2, X} \left[ \frac{1}{\sqrt{n}} \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \dot{\Sigma}_X \right\|_F \gtrsim B_{\mu,\epsilon} + \gamma_7 \gamma_2 K_\mu L_f^2 \log^3(nN_{X,\epsilon}) \sqrt{\frac{n}{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \right] + \Pr_{Y_{X,\epsilon}|X} [\mathcal{E}_2^C] \quad (\text{see (57)}) \\
& \lesssim (nN_{X,\epsilon})^{(1-K_\mu^2 \gamma_2) \log(nN_{X,\epsilon})} + n^{2-\log \gamma_7} + N^{1-\log \gamma_3}, \quad (\text{see Lemma 4 and (41)})
\end{aligned} \tag{53}$$

under (39) and (52). Lastly, we reintroduce  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  by invoking Lemma 1: It holds that

$$\begin{aligned}
& \frac{1}{\sqrt{n}} \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \dot{\Sigma}_X \right\|_F \\
& \leq \frac{1}{\sqrt{n}} \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \ddot{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F + \frac{1}{\sqrt{n}} \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \dot{\Sigma}_X \right\|_F \quad (\text{triangle inequality}) \\
& \lesssim \epsilon L_f H_f Q_{X,\epsilon}^{\frac{1}{2}} n + B_{\mu,\epsilon} + \gamma_7 \gamma_2 K_\mu L_f^2 \log^3(nN_{X,\epsilon}) \sqrt{\frac{n}{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \quad (\text{see Lemma 1}) \\
& = \epsilon L_f H_f \sqrt{\gamma_2 \log N_{X,\epsilon}} n + B_{\mu,\epsilon} + \gamma_7 \gamma_2 K_\mu L_f^2 \log^3(nN_{X,\epsilon}) \sqrt{\frac{n}{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \quad (\text{choice of } Q_{X,\epsilon} \text{ in (36)}),
\end{aligned} \tag{54}$$

with a failure probability of the order of

$$\begin{aligned}
& N_{X,\epsilon}^{1-K_\mu \gamma_2} + \left( (nN_{X,\epsilon})^{(1-K_\mu^2 \gamma_2) \log(nN_{X,\epsilon})} + n^{2-\log \gamma_7} + N^{1-\log \gamma_3} \right) \quad (\text{see (36) and (53)}) \\
& \lesssim N_{X,\epsilon}^{1-\min[K_\mu, K_\mu^2] \gamma_2} + n^{2-\log \gamma_7} + N^{1-\log \gamma_3},
\end{aligned} \tag{55}$$

and under (39) and (52). This completes the proof of Theorem 2 after taking  $\gamma_3 = \gamma_2 K_\mu$  and  $\gamma_7 = \gamma_2$ .

## References

- [1] J. F. Traub and H. Wozniakowski. A general theory of optimal algorithms. Technical report, Academic Press New York, 1980.
- [2] A. Cohen, I. Daubechies, R. DeVore, G. Kerkyacharian, and D. Picard. Capturing ridge functions in high dimensions from point queries. *Constructive Approximation*, 35(2):225–243, 2012.
- [3] M. Fornasier, K. Schnass, and J. Vybiral. Learning functions of few arbitrary linear parameters in high dimensions. *Foundations of Computational Mathematics*, 12(2):229–262, 2012.
- [4] J. Haupt, R. M Castro, and R. Nowak. Distilled sensing: Adaptive sampling for sparse detection and estimation. *IEEE Transactions on Information Theory*, 57(9):6222–6235, 2011.
- [5] P. G. Constantine. *Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies*. SIAM Spotlights. Society for Industrial and Applied Mathematics, 2015.
- [6] H. Tyagi and V. Cevher. Learning non-parametric basis independent models from point queries via low-rank methods. *Applied and Computational Harmonic Analysis*, 37(3):389–412, 2014.
- [7] K. Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.
- [8] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.
- [9] P. G. Constantine, A. Eftekhari, and R. Ward. A near-stationary subspace for ridge approximation. *arXiv preprint arXiv:1606.01929*, 2016.
- [10] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.
- [11] P. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.
- [12] A. Pinkus. *Ridge Functions*. Cambridge Tracts in Mathematics. Cambridge University Press, 2015.
- [13] R. DeVore, G. Petrova, and P. Wojtaszczyk. Approximation of functions of few variables in high dimensions. *Constructive Approximation*, 33(1):125–143, 2011.
- [14] C. J. Stone. Additive regression and other nonparametric models. *The annals of Statistics*, pages 689–705, 1985.
- [15] S. Gaïffas and G. Lecué. Optimal rates and adaptation in the single-index model using aggregation. *Electronic Journal of Statistics*, 1:538–573, 2007.
- [16] A. B. Juditsky, O. V. Lepski, and A. B. Tsybakov. Nonparametric estimation of composite functions. *The Annals of Statistics*, pages 1360–1404, 2009.
- [17] G. K. Golubev. Asymptotic minimax estimation of regression in the additive model. *Problemy peredachi informatsii*, 28(2):3–15, 1992.
- [18] E. Novak and H. Woźniakowski. *Tractability of Multivariate Problems: Standard information for functionals*, volume 12. European Mathematical Society, 2010.
- [19] E. J. Candes. *Ridgelets: Theory and applications*. PhD thesis, Stanford University, 1998.
- [20] S. Keiper. Analysis of generalized ridge functions in high dimensions. In *International Conference on Sampling Theory and Applications (SampTA)*, pages 259–263. IEEE, 2015.



- [21] I. Bogunovic, V. Cevher, J. Haupt, and J. Scarlett. Active learning of self-concordant like multi-index functions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2189–2193. IEEE, 2015.
- [22] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1990.
- [23] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823, 1981.
- [24] P. J. Huber. Projection pursuit. *The annals of Statistics*, pages 435–475, 1985.
- [25] D. L. Donoho and I. M. Johnstone. Projection-based approximation and a duality with kernel methods. *The Annals of Statistics*, pages 58–106, 1989.
- [26] A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999.
- [27] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.
- [28] F. Vivarelli and C. K. I. Williams. Discovering hidden features with Gaussian processes regression. *Advances in Neural Information Processing Systems*, pages 613–619, 1999.
- [29] R. Tripathy, I. Bilonis, and M. Gonzalez. Gaussian processes with built-in dimensionality reduction: Applications to high-dimensional uncertainty propagation. *Journal of Computational Physics*, 321:191–223, 2016.
- [30] K. C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.
- [31] X. Yin and B. Li. Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *The Annals of Statistics*, pages 3392–3416, 2011.
- [32] R. D. Cook. Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *Proceedings of the section on Physical and Engineering Sciences*, pages 18–25. American Statistical Association Alexandria, VA, 1994.
- [33] Y. Xia, H. Tong, W. K. Li, and L. X. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410, 2002.
- [34] K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *The Journal of Machine Learning Research*, 5:73–99, 2004.
- [35] A. M. Samarov. Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association*, 88(423):836–847, 1993.
- [36] M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny. Structure adaptive approach for dimension reduction. *The Annals of Statistics*, 29(6):1537–1566, 2001.
- [37] T. T. Cai and A. Zhang. ROP: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138, 2015.
- [38] Y. Chen, Y. Chi, and A. J. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.
- [39] G. Dasarthy, P. Shah, B. Narayan Bhaskar, and R. D. Nowak. Sketching sparse matrices, covariances, and graphs via tensor products. *IEEE Transactions on Information Theory*, 61(3):1373–1388, 2015.
- [40] M. Kolar and E. P. Xing. Consistent covariance selection from data with missing values. In *Proceedings of the International Conference on Machine Learning (ICML-12)*, pages 551–558, 2012.

- [41] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing  $\ell_1$ -penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.
- [42] M. Azizyan, A. Krishnamurthy, and A. Singh. Extreme compressive sampling for covariance estimation. *arXiv preprint arXiv:1506.00898*, 2015.
- [43] A. Krishnamurthy, M. Azizyan, and A. Singh. Subspace learning from extremely compressed measurements. *arXiv preprint arXiv:1404.0751*, 2014.
- [44] F. P. Anaraki and S. Hughes. Memory and computation efficient PCA via very sparse random projections. In *Proceedings of the International Conference on Machine Learning (ICML-14)*, pages 1341–1349, 2014.
- [45] F. Pourkamali-Anaraki. Estimation of the sample covariance matrix from compressive measurements. *arXiv preprint arXiv:1512.08887*, 2015.
- [46] P. L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2011.
- [47] A. Gonen, D. Rosenbaum, Y. Eldar, and S. Shalev-Shwartz. The sample complexity of subspace learning with partial information. *arXiv preprint arXiv:1402.4844*, 2014.
- [48] D. R. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization*, 21(4):345–383, 2001.
- [49] R. R. Barton. Metamodels for simulation input-output relations. In *Proceedings of the 24th conference on Winter simulation*, pages 289–299. ACM, 1992.
- [50] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statistical Science*, pages 409–423, 1989.
- [51] M. C. Kennedy and A. O’Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.
- [52] B. F. Logan and L. A. Shepp. Optimal reconstruction of a function from its projections. *Duke Mathematics Journal*, 42(4):645–659, 1975.
- [53] S. Shan and G. G. Wang. Survey of modeling and optimization strategies to solve high-dimensional design problems with computationally-expensive black-box functions. *Structural and Multidisciplinary Optimization*, 41(2):219–241, 2010.
- [54] A. Cohen, R. DeVore, and C. Schwab. Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs. *Foundations of Computational Mathematics*, 10(6):615–646, 2010.
- [55] B. Recht. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, 12:3413–3430, 2011.
- [56] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.
- [57] M. B. Eftekhari, A. Wakin, and R. A. Ward. Mc<sup>2</sup>: A two-phase algorithm for leveraged matrix completion. *arXiv preprint arXiv:1609.01795*, 2016.
- [58] A. Eftekhari, L. Balzano, and M. B. Wakin. What to expect when you are expecting on the Grassmannian. *arXiv preprint arXiv:1611.07216*, 2016.
- [59] T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *Annual IEEE Symposium on Foundations of Computer Science (FOCS’06)*, pages 143–152. IEEE, 2006.
- [60] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

- [61] E. Liberty, F. Woolfe, P. G. Martinsson, V. Rokhlin, and M. Tygert. Randomized algorithms for the low-rank approximation of matrices. *Proceedings of the National Academy of Sciences*, 104(51):20167–20172, 2007.
- [62] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.
- [63] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk. Beyond Nyquist: Efficient sampling of sparse bandlimited signals. *IEEE Transactions on Information Theory*, 56(1):520–544, 2010.

## A Toolbox

In this section, we list a few results that are repeatedly used in the rest of appendices. Recall the following inequalities for a random variable  $z$  and event  $\mathcal{A}$  (with complement  $\mathcal{A}^C$ ):<sup>5</sup>

$$\begin{aligned}\mathbb{E}_z^p[z] &\leq \mathbb{E}_{z|\mathcal{A}}^p[z] + \sup |z| \cdot \left(\Pr_z[\mathcal{A}^C]\right)^{\frac{1}{p}}, \quad (\text{if } \sup |z| < \infty), \\ \Pr_z[z > z_0] &\leq \Pr_{z|\mathcal{A}}[z > z_0] + \Pr_z[\mathcal{A}^C], \quad \forall z_0.\end{aligned}\tag{57}$$

We also recall the Bernstein inequality [62].

**Proposition 2. (Bernstein inequality)** *Let  $\{A_i\}_i$  be a finite sequence of zero-mean independent random matrices, and set*

$$b := \max_i \|A_i\|_F, \tag{58}$$

$$\sigma^2 := \sum_i \mathbb{E} \|A_i\|_F^2. \tag{59}$$

*Then, for  $\gamma \geq 1$  and except with a probability of at most  $e^{-\gamma}$ , it holds that*

$$\left\| \sum_i A_i \right\|_F \lesssim \gamma \cdot \max[b, \sigma]. \tag{60}$$

## B Proof of Proposition 1

Recalling the definition of  $\dot{\Sigma}_X$  from (11), we write that

$$\begin{aligned}\mathbb{E}_X [\dot{\Sigma}_X] &= \frac{1}{N} \sum_{x \in X} \mathbb{E}_X [\nabla f(x) \nabla f(x)^*] \quad (\text{see (11)}) \\ &= \mathbb{E}_x [\nabla f(x) \nabla f(x)^*] \quad (\#X = N) \\ &= \Sigma_\mu, \quad (\text{see (10)})\end{aligned}\tag{61}$$

---

<sup>5</sup>To see why the first inequality holds, note that

$$\begin{aligned}\mathbb{E}_z^p[z] &= \mathbb{E}_z^p[z \cdot 1_{\mathcal{A}}(z) + z \cdot 1_{\mathcal{A}^C}(z)] \\ &\leq \mathbb{E}_z^p[z \cdot 1_{\mathcal{A}}(z)] + \sup |z| \cdot \mathbb{E}^p[z \cdot 1_{\mathcal{A}^C}(z)], \quad (\text{triangle inequality})\end{aligned}$$

where  $1_{\mathcal{A}}(\cdot)$  is the indicator function for the event  $\mathcal{A}$ . It is easily verified that

$$\mathbb{E}_z^p[z \cdot 1_{\mathcal{A}}(z)] \leq \mathbb{E}_{z|\mathcal{A}}^p[z], \quad \mathbb{E}_z^p[z \cdot 1_{\mathcal{A}^C}(z)] \leq \sup |z| \cdot \Pr_z[\mathcal{A}^C]^{\frac{1}{p}}, \tag{56}$$

from which (57) follows immediately.

which proves the first claim. To control the deviation about the mean, we will invoke the standard Bernstein inequality, recorded in Proposition 2 for the reader's convenience. Note that

$$\begin{aligned}
\dot{\Sigma}_X - \Sigma_\mu &= \dot{\Sigma}_X - \mathbb{E}_X [\dot{\Sigma}_X] \\
&= \frac{1}{N} \sum_{x \in X} \nabla f(x) \nabla f(x)^* - \mathbb{E}_x [\nabla f(x) \nabla f(x)^*] \\
&=: \sum_{x \in X} A_x,
\end{aligned} \tag{62}$$

where  $\{A_x\}_x \subset \mathbb{R}^{n \times n}$  are independent and zero-mean random matrices. To apply the Bernstein inequality (Proposition 2), we compute the parameters  $b$  and  $\sigma$  next:

$$\begin{aligned}
b &= \max_{x \in X} \|A_x\|_F \\
&= \frac{1}{N} \max_{x \in X} \|\nabla f(x) \nabla f(x)^* - \mathbb{E}_x [\nabla f(x) \nabla f(x)^*]\|_F \quad (\text{see (62)}) \\
&\leq \frac{1}{N} \max_{x \in X} \|\nabla f(x) \nabla f(x)^*\|_F + \frac{1}{N} \mathbb{E}_x \|\nabla f(x) \nabla f(x)^*\|_F \quad (\text{triangle and Jensen's inequalities}) \\
&\leq \frac{2}{N} \sup_{x \in \mathbb{D}} \|\nabla f(x) \nabla f(x)^*\|_F \\
&= \frac{2}{N} \sup_{x \in \mathbb{D}} \|\nabla f(x)\|_2^2 \\
&= \frac{2L_f^2}{N}, \quad (\text{see (8)})
\end{aligned}$$

$$\begin{aligned}
\sigma^2 &= \sum_{x \in X} \mathbb{E}_x \|A_x\|_F^2 \\
&= \frac{1}{N} \mathbb{E}_x \|\nabla f(x) \nabla f(x)^* - \mathbb{E}_x [\nabla f(x) \nabla f(x)^*]\|_F^2 \quad (\text{see (62) and } \#X = N) \\
&\leq \frac{1}{N} \mathbb{E}_x \|\nabla f(x) \nabla f(x)^*\|_F^2 \quad (\mathbb{E}\|Z - \mathbb{E}[Z]\|_F^2 \leq \mathbb{E}\|Z\|_F^2 \text{ for a random matrix } Z) \\
&= \frac{1}{N} \mathbb{E}_x \|\nabla f(x)\|_2^4 \\
&\leq \frac{L_f^4}{N}, \quad (\text{see (8)})
\end{aligned}$$

$$\max[b, \sigma] \leq \frac{2L_f^2}{\sqrt{N}}. \tag{63}$$

Therefore, for  $\gamma_4 \geq 1$  and except with a probability of at most  $e^{-\gamma_4}$ , Proposition 2 dictates that

$$\begin{aligned}
\left\| \dot{\Sigma}_X - \Sigma_\mu \right\|_F &= \left\| \sum_{x \in X} A_x \right\|_F \quad (\text{see (62)}) \\
&\lesssim \gamma_4 \cdot \max[b, \sigma] \\
&\lesssim \gamma_4 \cdot \frac{L_f^2}{\sqrt{N}},
\end{aligned}$$

which completes the proof of Proposition 1 when we take  $\gamma_4 = \log n$ .

## C Uniform Measure Satisfies Assumption 1

We verify in this appendix that the uniform probability measure on  $\mathbb{D}$  satisfies Assumption 1. Fix arbitrary  $\epsilon > 0$  and  $x$  in the  $\epsilon$ -interior of  $\mathbb{D} \subseteq \mathbb{R}^n$ , namely  $x \in \mathbb{D}_\epsilon$ , assuming that  $\mathbb{D}_\epsilon \neq \emptyset$ . The conditional measure in the neighborhood  $\mathbb{B}_{x,\epsilon}$  too is uniform, so that  $y|x \sim \text{uniform}(\mathbb{B}_{x,\epsilon})$ . Then, for fixed  $v \in \mathbb{R}^n$  with  $\|v\|_2 = 1$ , observe that

$$\|P_{x,y}v\|_2^2 \sim \text{beta}\left(\frac{1}{2}, \frac{n-1}{2}\right). \quad (64)$$

To study the tail bound of the random variable  $\|P_{x,y}v\|_2^2$ , we proceed as follows. We fix  $\gamma_1 > 0$  and, recalling the moments of the beta distribution, write that

$$\begin{aligned} \Pr\left[\|P_{x,y}v\|_2^2 > \frac{\gamma_1}{n}\right] &= \Pr\left[\|P_{x,y}v\|_2^{2\lambda} > \left(\frac{\gamma_1}{n}\right)^\lambda\right] \quad (\lambda > 0) \\ &\leq \left(\frac{\gamma_1}{n}\right)^{-\lambda} \mathbb{E}\left[\|P_{x,y}v\|_2^{2\lambda}\right] \quad (\text{Markov's inequality}) \\ &= \left(\frac{\gamma_1}{n}\right)^{-\lambda} \frac{B\left(\lambda + \frac{1}{2}, \frac{n-1}{2}\right)}{B\left(\frac{1}{2}, \frac{n-1}{2}\right)}, \end{aligned} \quad (65)$$

where

$$B(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \quad (66)$$

is the beta function. Above,  $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$  is the usual gamma function. In order to choose  $\lambda$  above, we rewrite (65) as

$$\begin{aligned} \Pr\left[\|P_{x,y}v\|_2^2 > \frac{\gamma_1}{n}\right] &\leq e^{-\lambda \log\left(\frac{\gamma_1}{n}\right) + \log\left(B\left(\lambda + \frac{1}{2}, \frac{n-1}{2}\right)\right) - \log\left(B\left(\frac{1}{2}, \frac{n-1}{2}\right)\right)} \\ &=: e^{l(\lambda)}. \end{aligned} \quad (67)$$

In order to minimize  $l(\cdot)$ , we compute its derivative:

$$\begin{aligned} l'(\lambda) &= -\log\left(\frac{\gamma_1}{n}\right) + \frac{d}{d\lambda} \log\left(B\left(\lambda + \frac{1}{2}, \frac{n-1}{2}\right)\right) - \frac{d}{d\lambda} \log\left(B\left(\frac{1}{2}, \frac{n-1}{2}\right)\right) \\ &= -\log\left(\frac{\gamma_1}{n}\right) + \frac{d}{d\lambda} \log\left(\Gamma\left(\lambda + \frac{1}{2}\right)\right) - \frac{d}{d\lambda} \log\left(\Gamma\left(\lambda + \frac{n}{2}\right)\right) \quad (\text{see (66)}) \\ &= -\log\left(\frac{\gamma_1}{n}\right) + \frac{\Gamma'\left(\lambda + \frac{1}{2}\right)}{\Gamma\left(\lambda + \frac{1}{2}\right)} - \frac{\Gamma'\left(\lambda + \frac{n}{2}\right)}{\Gamma\left(\lambda + \frac{n}{2}\right)} \\ &= -\log\left(\frac{\gamma_1}{n}\right) + \psi\left(\lambda + \frac{1}{2}\right) - \psi\left(\lambda + \frac{n}{2}\right), \end{aligned} \quad (68)$$

where  $\psi(a) = \frac{\Gamma'(a)}{\Gamma(a)}$  is the “digamma” function. It is well-known that  $\psi(a) \approx \log(a)$  for large  $a$  (see, for example, the corresponding Wikipedia entry). If  $n$  is sufficiently large and we take  $\lambda$  such that  $1 \ll \lambda \ll n$ , we have that

$$\begin{aligned} l'(\lambda) &= -\log\left(\frac{\gamma_1}{n}\right) + \psi\left(\lambda + \frac{1}{2}\right) - \psi\left(\lambda + \frac{n}{2}\right) \quad (\text{see (68)}) \\ &\approx -\log\left(\frac{\gamma_1}{n}\right) + \log \lambda - \log\left(\frac{n}{2}\right) \\ &= -\log\left(\frac{2\lambda}{\gamma_1}\right), \end{aligned} \quad (69)$$

thereby suggesting the choice of  $\lambda = \gamma_1/2$ . With this choice, we find that

$$\begin{aligned}
\Pr \left[ \|P_{x,y}v\|_2^2 > \frac{\gamma_1}{n} \right] &\leq \left( \frac{\gamma_1}{n} \right)^{-\frac{\gamma_1}{2}} \frac{B\left(\frac{\gamma_1+1}{2}, \frac{n-1}{2}\right)}{B\left(\frac{1}{2}, \frac{n-1}{2}\right)} \quad (\text{see (65)}) \\
&= \left( \frac{\gamma_1}{n} \right)^{-\frac{\gamma_1}{2}} \frac{\Gamma\left(\frac{\gamma_1+1}{2}\right) \Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{1}{2}\right) \Gamma\left(\frac{n+\gamma_1}{2}\right)} \quad (\text{see (66)}) \\
&\lesssim \left( \frac{\gamma_1}{n} \right)^{-\frac{\gamma_1}{2}} \frac{\left(\frac{\gamma_1+1}{2}\right)^{\frac{\gamma_1}{2}} e^{-\frac{\gamma_1+1}{2}} \left(\frac{n}{2}\right)^{\frac{n-1}{2}} e^{-\frac{n}{2}}}{\left(\frac{n+\gamma_1}{2}\right)^{\frac{n+\gamma_1-1}{2}} e^{-\frac{n+\gamma_1}{2}}} \quad \left( 1 < \frac{a^{\frac{1}{2}-a} e^a}{\sqrt{2\pi}} \Gamma(a) < e^{\frac{1}{12a}}, \forall a > 0 \right) \\
&\lesssim \left( \frac{n}{n+\gamma_1} \right)^{\frac{n+\gamma_1-1}{2}} \\
&\leq \left( \frac{n}{n+\gamma_1} \right)^{\frac{n-1}{2}} \quad (\gamma_1 > 0) \\
&= \left( 1 + \frac{\gamma_1}{n} \right)^{-\frac{n-1}{2}} \\
&\leq e^{-\frac{\gamma_1}{n} \cdot \frac{n-1}{2}} \quad (1+a \leq e^a) \\
&\leq e^{-\frac{\gamma_1}{2} + \frac{1}{2}}, \quad (\gamma_1 \leq n)
\end{aligned} \tag{70}$$

Therefore, Assumption 1 holds for the uniform probability measure with  $\epsilon_\mu = \infty$  and  $K_\mu = 1/2$ .

## D Estimating $\nabla f(x)$

For fixed  $x \in \mathbb{D}$ , by drawing samples from the neighborhood of  $x$  and then applying the method of finite differences, we may estimate  $\nabla f(x)$ . This is described below for the sake of completeness.

**Proposition 3.** Fix  $x \in \mathbb{D}$  and take  $\epsilon > 0$  small enough so that  $x$  belongs to  $\epsilon$ -interior of  $\mathbb{D}$ , namely  $x \in \mathbb{D}_\epsilon$ . Draw  $y$  from the conditional measure on the neighborhood  $\mathbb{B}_{x,\epsilon}$ , namely  $y|x \sim \mu_{x,\epsilon}$  (see (17)). For an integer  $N_{x,\epsilon}$ , let  $Y_{x,\epsilon} \subset \mathbb{B}_{x,\epsilon}$  contain  $N_{x,\epsilon}$  independent copies of  $y$ . Then, it holds that

$$\left\| \mathbb{E}_{Y_{x,\epsilon} | N_{x,\epsilon}, x} [\dot{\nabla}_{Y_{x,\epsilon}} f(x)] - \nabla f(x) \right\|_2 \leq B'_{\mu,\epsilon} L_f + \frac{\epsilon H_f n}{2}. \tag{71}$$

where

$$\dot{\nabla}_{Y_{x,\epsilon}} f(x) := \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} \frac{f(y) - f(x)}{\|y - x\|_2} \cdot \frac{y - x}{\|y - x\|_2} \in \mathbb{R}^n, \tag{72}$$

$$B'_{\mu,\epsilon} := n \cdot \sup_{x \in \mathbb{D}_\epsilon} \left\| \mathbb{E}_{y|x} [P_{x,y}] - \frac{I_n}{n} \right\|. \quad (y|x \sim \mu_{x,\epsilon}) \tag{73}$$

In particular, if  $\mu$  is the uniform probability measure on  $\mathbb{D}$ , then  $B_{\mu,\epsilon} = 0$ .

*Proof.* First, we replace  $\dot{\nabla}_{Y_{x,\epsilon}} f(x)$  with the simpler quantity  $\ddot{\nabla}_{Y_{x,\epsilon}} f(x)$ , defined as

$$\ddot{\nabla}_{Y_{x,\epsilon}} f(x) := \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} P_{x,y} \cdot \nabla f(x) \in \mathbb{R}^n, \tag{74}$$

where  $P_{x,y} \in \mathbb{R}^{n \times n}$  is the orthogonal projection onto the direction of  $y - x$ . By definition, the two quantities

are related as follows:

$$\begin{aligned}
& \left\| \dot{\nabla}_{Y_{x,\epsilon}} f(x) - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2 \\
&= \frac{n}{N_{x,\epsilon}} \left\| \sum_{y \in Y_{x,\epsilon}} \frac{y-x}{\|y-x\|_2^2} (f(y) - f(x) - (y-x)^* \nabla f(x)) \right\|_2 \quad \left( P_{x,y} = \frac{(y-x)(y-x)^*}{\|y-x\|_2^2} \right) \\
&\leq \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} \left\| \frac{y-x}{\|y-x\|_2^2} (f(y) - f(x) - (y-x)^* \nabla f(x)) \right\|_2 \quad (\text{triangle inequality}) \\
&= \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} \frac{|f(y) - f(x) - (y-x)^* \nabla f(x)|}{\|y-x\|_2} \\
&\leq n \cdot \sup_{y \in \mathbb{B}_{x,\epsilon}} \frac{|f(y) - f(x) - (y-x)^* \nabla f(x)|}{\|y-x\|_2} \quad (\#Y_{x,\epsilon} = N_{x,\epsilon}) \\
&\leq n \cdot \sup_{y \in \mathbb{B}_{x,\epsilon}} \frac{H_f \|y-x\|_2}{2} \quad (\text{Taylor's expansion and (9)}) \\
&\leq n \cdot \frac{H_f \cdot \epsilon}{2}. \quad (y \in \mathbb{B}_{x,\epsilon})
\end{aligned} \tag{75}$$

Loosely speaking then,  $\dot{\nabla}_{Y_{x,\epsilon}} f(x) \approx \ddot{\nabla}_{Y_{x,\epsilon}} f(x)$  and it therefore suffices to study the estimation bias of  $\ddot{\nabla}_{Y_{x,\epsilon}} f(x)$ . To that end, we simply note that

$$\begin{aligned}
\left\| \mathbb{E}_{Y_{x,\epsilon} | N_{x,\epsilon}, x} [\dot{\nabla}_{Y_{x,\epsilon}} f(x)] - \nabla f(x) \right\|_2 &= \left\| n \cdot \mathbb{E}_{y|x} [P_{x,y} \nabla f(x)] - \nabla f(x) \right\|_2 \quad (y|x \sim \mu_{x,\epsilon}) \\
&= \left\| n \cdot \mathbb{E}_{y|x} [P_{x,y}] \cdot \nabla f(x) - \nabla f(x) \right\|_2 \\
&\leq n \cdot \sup_{x \in \mathbb{D}} \left\| \mathbb{E}_{y|x} [P_{x,y}] - \frac{I_n}{n} \right\| \cdot \sup_{x \in \mathbb{D}} \|\nabla f(x)\|_2 \\
&=: B'_{\mu,\epsilon} \cdot L_f, \quad (\text{see (8)})
\end{aligned} \tag{76}$$

which, in turn, implies that

$$\begin{aligned}
& \left\| \mathbb{E}_{Y_{x,\epsilon} | N_{x,\epsilon}, x} [\dot{\nabla}_{Y_{x,\epsilon}} f(x)] - \nabla f(x) \right\|_2 \\
&\leq \left\| \mathbb{E}_{Y_{x,\epsilon} | N_{x,\epsilon}, x} [\dot{\nabla}_{Y_{x,\epsilon}} f(x) - \ddot{\nabla}_{Y_{x,\epsilon}} f(x)] \right\|_2 + \left\| \mathbb{E}_{Y_{x,\epsilon} | N_{x,\epsilon}, x} [\ddot{\nabla}_{Y_{x,\epsilon}} f(x)] - \nabla f(x) \right\|_2 \quad (\text{triangle inequality}) \\
&\leq \frac{n H_f \epsilon}{2} + B'_{\mu,\epsilon} L_f. \quad (\text{see (75) and (76)})
\end{aligned} \tag{77}$$

In particular, when  $\mu$  is the uniform probability measure on  $\mathbb{D}$ ,  $P_{x,y}$  is an isotropic random matrix (for fixed  $x \in \mathbb{D}$ ). Therefore,  $\mathbb{E}_{y|x} [P_{x,y}] = C \cdot I_n$  for some scalar  $C$ . To find  $C$ , we note that

$$\text{trace} [\mathbb{E}_{y|x} [P_{x,y}]] = \mathbb{E}_{y|x} [\text{trace} [P_{x,y}]] = 1 = C \cdot \text{trace} [I_n] = C \cdot n \implies C = \frac{1}{n},$$

where we used the fact that  $P_{x,y}$  is a rank-1 orthogonal projection. Consequently, when  $\mu$  is the uniform measure,  $B_{\mu,\epsilon} = 0$ . This completes the proof of Proposition 3.  $\square$

## E Proof of Lemma 1

We only verify the second claim, as the other proof is similar. Conditioned on the event  $\mathcal{E}_1$ , note that

$$\begin{aligned}
\left\| \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2 &\leq \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} \|P_{x,y} \cdot \nabla f(x)\|_2 \quad (\text{see (32)}) \\
&\leq n \cdot \max_{x \in X} \max_{y \in Y_{x,\epsilon}} \|P_{x,y} \cdot \nabla f(x)\|_2 \quad (\#Y_{x,\epsilon} = N_{x,\epsilon}) \\
&\leq n \cdot \sqrt{\frac{Q_{X,\epsilon} L_f^2}{n}}. \quad (\text{see (34)})
\end{aligned} \tag{78}$$

Using the inequality  $\|aa^* - bb^*\|_2 \leq \|a - b\|(\|a\|_2 + \|b\|_2)$  for any  $a, b \in \mathbb{R}^n$  in the third line below, it follows that

$$\begin{aligned}
&\frac{1}{N} \left\| \sum_{N_{x,\epsilon} > N_{X,\min,\epsilon}} \dot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \dot{\nabla}_{Y_{x,\epsilon}} f(x)^* - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \ddot{\nabla}_{Y_{x,\epsilon}} f(x)^* \right\|_F \\
&\leq \frac{1}{N} \sum_{N_{x,\epsilon} > N_{X,\min,\epsilon}} \left\| \dot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \dot{\nabla}_{Y_{x,\epsilon}} f(x)^* - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \ddot{\nabla}_{Y_{x,\epsilon}} f(x)^* \right\|_F \quad (\text{triangle inequality}) \\
&\leq \frac{1}{N} \sum_{N_{x,\epsilon} > N_{X,\min,\epsilon}} \left\| \dot{\nabla}_{Y_{x,\epsilon}} f(x) - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2 \left( \left\| \dot{\nabla} f(x) \right\|_2 + \left\| \ddot{\nabla} f(x) \right\|_2 \right) \\
&\leq \max_{x \in X} \left\| \dot{\nabla}_{Y_{x,\epsilon}} f(x) - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2 \left( \max_{x \in X} \left\| \dot{\nabla} f(x) \right\|_2 + \max_{x \in X} \left\| \ddot{\nabla} f(x) \right\|_2 \right) \quad (\#X = N) \\
&\leq \max_{x \in X} \left\| \dot{\nabla}_{Y_{x,\epsilon}} f(x) - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2 \left( \max_{x \in X} \left\| \dot{\nabla} f(x) - \ddot{\nabla} f(x) \right\|_2 + 2 \max_{x \in X} \left\| \ddot{\nabla} f(x) \right\|_2 \right) \quad (\text{triangle inequality}) \\
&\leq \frac{\epsilon H_f n}{2} \left( \frac{\epsilon L_f n}{2} + 2 \sqrt{Q_{X,\epsilon} L_f^2 n} \right) \quad (\text{see (75) and (78)}) \\
&\lesssim \epsilon L_f H_f Q_{X,\epsilon}^{\frac{1}{2}} n^{\frac{3}{2}}, \quad \left( \text{when } \epsilon \lesssim \sqrt{\frac{Q_{X,\epsilon}}{n}} \right)
\end{aligned} \tag{79}$$

which, in turn, immediately implies that

$$\begin{aligned}
&\frac{1}{N} \left| \sum_{N_{x,\epsilon} > N_{X,\min,\epsilon}} \left\| \dot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2^2 - \left\| \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2^2 \right| \\
&= \frac{1}{N} \left| \sum_{N_{x,\epsilon} > N_{X,\min,\epsilon}} \text{trace} \left[ \dot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \dot{\nabla}_{Y_{x,\epsilon}} f(x)^* - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \ddot{\nabla}_{Y_{x,\epsilon}} f(x)^* \right] \right| \\
&\leq \frac{\sqrt{n}}{N} \left\| \sum_{N_{x,\epsilon} > N_{X,\min,\epsilon}} \dot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \dot{\nabla}_{Y_{x,\epsilon}} f(x)^* - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \ddot{\nabla}_{Y_{x,\epsilon}} f(x)^* \right\|_F \\
&\leq \sqrt{n} \cdot \epsilon L_f H_f Q_{X,\epsilon}^{\frac{1}{2}} n^{\frac{3}{2}}, \quad (\text{see (79)})
\end{aligned} \tag{80}$$



where the third line above uses the fact that  $|\text{trace}(A)| \leq \sqrt{n}\|A\|_F$  for any  $A \in \mathbb{R}^{n \times n}$ . Recall the definitions of  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  and  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  in (6) and (31), respectively. Then, by combining (79) and (80), it follows that

$$\begin{aligned}
& \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \ddot{\Sigma}_{X,Y_{X,\epsilon}} \right\|_F \\
& \leq \frac{1}{N} \left\| \sum_{N_{x,\epsilon} \geq N_{X,\min,\epsilon}} \dot{\nabla}_{Y_{x,\epsilon}} f(x) \dot{\nabla}_{Y_{x,\epsilon}} f(x)^* - \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \ddot{\nabla}_{Y_{x,\epsilon}} f(x)^* \right\|_F \\
& \quad + \frac{1}{N} \left| \sum_{N_{x,\epsilon} > N_{X,\min,\epsilon}} \left\| \dot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2^2 - \left\| \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \right\|_2^2 \right| \cdot \frac{\|I_n\|_F}{n} \quad (\text{see (6) and (31)}) \\
& \lesssim \epsilon L_f H_f Q_{X,\epsilon}^{\frac{1}{2}} n^{\frac{3}{2}}. \quad (\text{see (79) and (80)})
\end{aligned} \tag{81}$$

This completes the proof of Lemma 1.

## F Proof of Lemma 2

Our objective is to establish that, given  $X$  and neighborhood radius  $\epsilon$ , each  $x \in X$  has many neighbors in  $Y_{X,\epsilon}$  provided that  $N_{X,\epsilon} = \#Y_{X,\epsilon}$  is sufficiently large. To that end, we proceed as follows. Recall that  $\mu_{X,\epsilon}$  is the conditional distribution on the  $\epsilon$ -neighborhood of the point cloud  $X$  (see (15)). With  $y \sim \mu_{X,\epsilon}$  and for fixed  $x \in X$ , observe that  $y$  belongs to the  $\epsilon$ -neighborhood of  $x$  (namely,  $y \in \mathbb{B}_{x,\epsilon}$ ) with the following probability:

$$\Pr_{y|x} [y \in \mathbb{B}_{x,\epsilon}] = \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})}. \tag{82}$$

Equivalently, the indicator function  $1_{y \in \mathbb{B}_{x,\epsilon}}$  follows a Bernoulli distribution:

$$1_{y \in Y_{x,\epsilon}} | x \sim \text{Bernoulli} \left( \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} \right). \tag{83}$$

Then,

$$\mathbb{E}_{Y_{X,\epsilon}|X} [N_{x,\epsilon}] = \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} \cdot \#Y_{X,\epsilon} = \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} \cdot N_{X,\epsilon}, \tag{84}$$

and, to investigate the concentration of  $N_{x,\epsilon}$  about its expectation, we write that

$$\begin{aligned}
N_{x,\epsilon} - \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} \cdot N_{X,\epsilon} &= N_{x,\epsilon} - \mathbb{E}_{Y_{X,\epsilon}|X} [N_{x,\epsilon}] \quad (\text{see (84)}) \\
&= \sum_{y \in Y_{X,\epsilon}} (1_{y \in \mathbb{B}_{x,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|X} [1_{y \in \mathbb{B}_{x,\epsilon}}]) \\
&= \sum_{y \in Y_{X,\epsilon}} \left( 1_{y \in \mathbb{B}_{x,\epsilon}} - \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} \right) \\
&=: \sum_{y \in Y_{X,\epsilon}} a_y,
\end{aligned} \tag{85}$$

where  $\{a_y\}_y$  are independent zero-mean random variables (for fixed  $x \in X$ ). In order to apply the Bernstein's inequality (Proposition 2) to the last line of (85), we write that

$$\begin{aligned}
b &= \max_y |a_y| \\
&= \max_y \left| 1_{y \in \mathbb{B}_{x,\epsilon}} - \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} \right| \quad (\text{see (85)}) \\
&\leq 1,
\end{aligned} \tag{86}$$

$$\begin{aligned}
\sigma^2 &= \sum_{y \in Y_{X,\epsilon}} \mathbb{E}_{Y_{x,\epsilon}|x} [a_y^2] \\
&= \sum_{y \in Y_{X,\epsilon}} \mathbb{E}_{Y_{x,\epsilon}|x} \left[ \left( 1_{y \in \mathbb{B}_{x,\epsilon}} - \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} \right)^2 \right] \quad (\text{see (85)}) \\
&= \sum_{y \in Y_{X,\epsilon}} \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} \left( 1 - \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} \right) \quad (\text{see (83)}) \\
&\leq \sum_{y \in Y_{X,\epsilon}} \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} = \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} \cdot N_{X,\epsilon}, \quad (\#Y_{x,\epsilon} = N_{x,\epsilon}) \tag{87}
\end{aligned}$$

$$\max[b, \sigma] = \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} N_{X,\epsilon} \cdot \left( \text{if } N_{X,\epsilon} \geq \frac{\mu(\mathbb{B}_{X,\epsilon})}{\mu(\mathbb{B}_{x,\epsilon})} \right) \tag{88}$$

From Proposition 2, then, it follows that

$$\begin{aligned}
\left| N_{x,\epsilon} - \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} N_{X,\epsilon} \right| &\lesssim \gamma_5 \cdot \max[b, \sigma] \\
&= \gamma_5 \cdot \sqrt{\frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} N_{X,\epsilon}}, \quad (\text{see (88)}) \tag{89}
\end{aligned}$$

for  $\gamma_5 \geq 1$  and except with a probability of at most  $e^{-\gamma_5}$ . Recall that  $\#X = N$ . Then, an application of the union bound with the choice of  $\gamma_5 = \gamma_3 \log N$  (with  $\gamma_3 \geq 1$ ) yields that

$$\max_{x \in X} \left| N_{x,\epsilon} - \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} N_{X,\epsilon} \right| \lesssim \gamma_3 \log N \cdot \sqrt{\frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} N_{X,\epsilon}}, \tag{90}$$

except with a probability of at most  $N e^{-\gamma_3 \log N} = N^{1-\log \gamma_3}$ . For the bound above to hold, we assume that  $N_{X,\epsilon}$  is sufficiently large (so that the requirement in (88) hold for every  $x \in X$ ). In fact, if

$$N_{X,\epsilon} \gtrsim \frac{\gamma_3^2 \log^2 N \cdot \mu(\mathbb{B}_{X,\epsilon})}{\min_{x \in X} \mu(\mathbb{B}_{x,\epsilon})}, \tag{91}$$

then (90) readily yields that

$$\frac{1}{2} \cdot \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} N_{X,\epsilon} \leq N_{x,\epsilon} \leq \frac{3}{2} \cdot \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} N_{X,\epsilon}, \quad \forall x \in X, \tag{92}$$

except with a probability of at most  $N^{1-\log \gamma_3}$ . This completes the proof of Lemma 2.

## G Proof of Lemma 3

Throughout,  $X$  and  $\epsilon \in (0, \epsilon_{\mu,X}]$  are fixed, and we further assume that the event  $\mathcal{E}_2$  holds (see (40)). For now, suppose in addition that the neighborhood structure  $\overline{N}_{X,\epsilon} := \{N_{x,\epsilon}\}_{x \in X}$  is fixed too. Recalling the definition of  $\ddot{\nabla}_{Y_{x,\epsilon}} f(\cdot)$  from (32), we first set

$$\mathbb{R}^{n \times n} \ni \ddot{\Sigma}_{X,Y_{X,\epsilon}} := \frac{1}{N} \sum_{x \in X} \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \ddot{\nabla}_{Y_{x,\epsilon}} f(x)^*, \tag{93}$$

for short, and then separate the “diagonal” and “off-diagonal” components of the expectation of  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  as follows:

$$\begin{aligned}
& \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] \\
&= \frac{1}{N} \cdot \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} \left[ \sum_{x \in X} \ddot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \ddot{\nabla}_{Y_{x,\epsilon}} f(x)^* \right] \quad (\text{see (93)}) \\
&= \frac{n^2}{N} \cdot \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} \left[ \sum_{x \in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y,y' \in Y_{x,\epsilon}} P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y'} \right] \quad (\text{see (32)}) \\
&= \frac{n^2}{N} \cdot \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} \left[ \sum_{x \in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y \in Y_{x,\epsilon}} P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y} \right] \\
&\quad + \frac{n^2}{N} \cdot \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} \left[ \sum_{x \in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y,y' \in Y_{x,\epsilon}} 1_{y \neq y'} \cdot P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y'} \right] \\
&= \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y \in Y_{x,\epsilon}} \mathbb{E}_{y|x} [P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}] \\
&\quad + \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y,y' \in Y_{x,\epsilon}} \mathbb{E}_{y,y'|x} [1_{y \neq y'} \cdot P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y'}] \quad (y, y' \sim \mu_{x,\epsilon}) \\
&= \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,\epsilon}} \cdot \mathbb{E}_{y|x} [P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}] \\
&\quad + \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y,y' \in Y_{x,\epsilon}} 1_{y \neq y'} \cdot \mathbb{E}_{y|x} [P_{x,y} \nabla f(x)] \cdot \mathbb{E}_{y'|x} [\nabla f(x)^* P_{x,y'}]. \tag{94}
\end{aligned}$$

The last line above uses the fact that distinct elements of  $Y_{x,\epsilon}$  are statistically independent. We next replace both the diagonal and off-diagonal components (namely, the first and second sums in the last line above) with simpler expressions. We approximate the diagonal term with another sum as follows:

$$\begin{aligned}
& \left\| \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,\epsilon}} \mathbb{E}_{y|x} [P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}] - \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,\epsilon}} \left( \frac{2 \nabla f(x) \nabla f(x)^*}{n(n+2)} + \frac{\|\nabla f(x)\|_2^2}{n(n+2)} \cdot I_n \right) \right\|_F \\
&\leq \frac{n^2}{\min_{x \in X} N_{x,\epsilon}} \cdot \sup_{x \in \mathbb{D}} \left\| \mathbb{E}_{y|x} [P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}] - \left( \frac{2 \nabla f(x) \nabla f(x)^*}{n(n+2)} + \frac{\|\nabla f(x)\|_2^2}{n(n+2)} \cdot I_n \right) \right\|_F \quad (\#X = N) \\
&=: \frac{B''_{\mu,\epsilon}}{\min_{x \in X} N_{x,\epsilon}} \\
&\leq \frac{B''_{\mu,\epsilon}}{N_{X,\min,\epsilon}}. \quad (\text{see (40)}) \tag{95}
\end{aligned}$$

To replace the off-diagonal term in the last line of (94), first recall the inequality

$$\|ab^* - cd^*\|_F \leq 2 \max[\|a - c\|_2, \|b - d\|_2] \cdot \max[\|b\|_2, \|c\|_2], \quad a, b, c, d \in \mathbb{R}^n, \tag{96}$$

and then note that

$$\begin{aligned}
& \left\| \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y, y' \in Y_{x,\epsilon}} 1_{y \neq y'} \cdot \mathbb{E}_{y|x} [P_{x,y} \nabla f(x)] \cdot \mathbb{E}_{y'|x} [\nabla f(x)^* P_{x,y'}] - \frac{1}{N} \sum_{x \in X} \frac{N_{x,\epsilon} - 1}{N_{x,\epsilon}} \nabla f(x) \nabla f(x)^* \right\|_F \\
&= \left\| \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y, y' \in Y_{x,\epsilon}} 1_{y \neq y'} \left( \mathbb{E}_{y|x} [P_{x,y} \nabla f(x)] \cdot \mathbb{E}_{y'|x} [\nabla f(x)^* P_{x,y'}] - \frac{\nabla f(x) \nabla f(x)^*}{n^2} \right) \right\|_F \quad (\#Y_{x,\epsilon} = N_{x,\epsilon}) \\
&\leq n^2 \max_{x \in X} \max_{y, y' \in Y_{x,\epsilon}} \left\| \mathbb{E}_{y|x} [P_{x,y} \nabla f(x)] \cdot \mathbb{E}_{y'|x} [\nabla f(x)^* P_{x,y'}] - \frac{\nabla f(x) \nabla f(x)^*}{n^2} \right\|_F \quad (\#X = N, \#Y_{X,\epsilon} = N_{x,\epsilon}) \\
&\leq 2n^2 \max_{x \in X} \left[ \left\| \mathbb{E}_{y|x} [P_{x,y} \nabla f(x)] - \frac{\nabla f(x)}{n} \right\|_2 \cdot \max \left[ \left\| \mathbb{E}_{y'|x} [P_{x,y'} \nabla f(x)^*] \right\|_2, \frac{\|\nabla f(x)\|_2}{n} \right] \right] \quad (\text{see (96)}) \\
&\leq 2n^2 \max_{x \in X} \left[ \left\| \mathbb{E}_{y|x} [P_{x,y} \nabla f(x)] - \frac{\nabla f(x)}{n} \right\|_2 \left( \left\| \mathbb{E}_{y'|x} [P_{x,y'} \nabla f(x)^*] - \frac{\nabla f(x)}{n} \right\|_2 + \frac{\|\nabla f(x)\|_2}{n} \right) \right] \\
&\leq 2n^2 \left( \frac{B'_{\mu,\epsilon}}{n} \cdot L_f \right) \left( \frac{B'_{\mu,\epsilon}}{n} \cdot L_f + \frac{L_f}{n} \right) \quad (\text{see (8) and (73)}) \\
&= 2B'_{\mu,\epsilon} (B'_{\mu,\epsilon} + 1) L_f^2. \tag{97}
\end{aligned}$$

We may now replace the diagonal and off diagonal components in the last line of (94) with simpler expressions while incurring a typically small error. More specifically, in light of (95) and (97), (94) now implies that

$$\begin{aligned}
& \left\| \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \frac{1}{N} \sum_{x \in X} \left( 1 + \frac{n-2}{N_{x,\epsilon}(n+2)} \right) \nabla f(x) \nabla f(x)^* - \frac{n}{N(n+2)} \sum_{x \in X} \frac{\|\nabla f(x)\|_2^2}{N_{x,\epsilon}} \cdot I_n \right\|_F \\
&= \left\| \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,\epsilon}} \left( \frac{2\nabla f(x) \nabla f(x)^*}{n(n+2)} + \frac{\|\nabla f(x)\|_2^2}{n(n+2)} \cdot I_n \right) \right. \\
&\quad \left. - \frac{1}{N} \sum_{x \in X} \frac{N_{x,\epsilon} - 1}{N_{x,\epsilon}} \nabla f(x) \nabla f(x)^* \right\|_F \\
&\leq \frac{B''_{\mu,\epsilon}}{N_{X,\min,\epsilon}} + 2B'_{\mu,\epsilon} (B'_{\mu,\epsilon} + 1) L_f^2. \quad (\text{see (95) and (97)}) \tag{98}
\end{aligned}$$

We can further simplify the first line of (98) by replacing  $N_{x,\epsilon}$  with  $N_{X,\min,\epsilon}$  as follows. By invoking (11) in the second line below, we note that

$$\begin{aligned}
& \left\| \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \left( 1 + \frac{n-2}{N_{X,\min,\epsilon}(n+2)} \right) \dot{\Sigma}_X - \frac{n}{N_{X,\min,\epsilon}(n+2)} \cdot \text{trace} [\dot{\Sigma}_X] \cdot I_n \right\|_F \\
&= \left\| \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \left( 1 + \frac{n-2}{N_{X,\min,\epsilon}(n+2)} \right) \frac{1}{N} \sum_x \nabla f(x) \nabla f(x)^* \right. \\
&\quad \left. - \frac{n}{N_{X,\min,\epsilon}(n+2)} \cdot \frac{1}{N} \sum_x \|\nabla f(x)\|_2^2 \cdot I_n \right\|_F \\
&\leq \left( \frac{B''_{\mu,\epsilon}}{N_{X,\min,\epsilon}} + 2B'_{\mu,\epsilon} (B'_{\mu,\epsilon} + 1) L_f^2 \right) + \max_{x \in X} \left| \frac{1}{N_{x,\epsilon}} - \frac{1}{N_{X,\min,\epsilon}} \right| \cdot \max_{x \in X} \|\nabla f(x)\|_2^2 \cdot (1 + \|I_n\|_F) \quad (\text{see (98)}) \\
&\leq \left( \frac{B''_{\mu,\epsilon}}{N_{X,\min,\epsilon}} + 2B'_{\mu,\epsilon} (B'_{\mu,\epsilon} + 1) L_f^2 \right) + \max_{x \in X} \left| \frac{1}{N_{x,\epsilon}} - \frac{1}{N_{X,\min,\epsilon}} \right| \cdot L_f^2 (1 + \sqrt{n}) \quad (\text{see (8)}) \\
&\leq \left( \frac{B''_{\mu,\epsilon}}{N_{X,\min,\epsilon}} + 2B'_{\mu,\epsilon} (B'_{\mu,\epsilon} + 1) L_f^2 \right) + \frac{L_f^2 (1 + \sqrt{n})}{N_{X,\min,\epsilon}} \quad (\text{see (40)}) \\
&=: \frac{1}{2} B_{\mu,\epsilon}. \tag{99}
\end{aligned}$$

Next, we replace  $\text{trace}[\ddot{\Sigma}_X]$  in the first line of (99) with  $\text{trace}[\ddot{\Sigma}_{X,Y_{X,\epsilon}}]$ . To that end, we first notice the following consequence of (99):

$$\begin{aligned}
& \left| \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} [\text{trace} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}]] - \left( 1 + \frac{n-2}{N_{X,\min,\epsilon}(n+2)} \right) \text{trace} [\dot{\Sigma}_X] - \frac{n^2}{N_{X,\min,\epsilon}(n+2)} \cdot \text{trace} [\dot{\Sigma}_X] \right| \\
&= \left| \text{trace} \left[ \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \left( 1 + \frac{n-2}{N_{X,\min,\epsilon}(n+2)} \right) \dot{\Sigma}_X - \frac{n}{N_{X,\min,\epsilon}(n+2)} \cdot \text{trace} [\dot{\Sigma}_X] \cdot I_n \right] \right| \\
&\leq \sqrt{n} \left\| \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \left( 1 + \frac{n-2}{N_{X,\min,\epsilon}(n+2)} \right) \dot{\Sigma}_X - \frac{n}{N_{X,\min,\epsilon}(n+2)} \cdot \text{trace} [\dot{\Sigma}_X] \cdot I_n \right\|_F \\
&\leq \frac{\sqrt{n}}{2} B_{\mu,\epsilon}, \quad (\text{see (99)})
\end{aligned} \tag{100}$$

where the second line uses the fact that  $\text{trace}[I_n] = n$ . Also, the third line follows from the inequality  $|\text{trace}[A]| \leq \sqrt{n}\|A\|_F$  for arbitrary matrix  $A \in \mathbb{R}^{n \times n}$ . After rearranging, (100) immediately implies that

$$\begin{aligned}
& \left| \left( 1 + \frac{n^2+n-2}{N_{X,\min,\epsilon}(n+2)} \right)^{-1} \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} [\text{trace} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}]] - \text{trace} [\dot{\Sigma}_X] \right| \\
&\leq \left( 1 + \frac{n^2+n-2}{N_{X,\min,\epsilon}(n+2)} \right)^{-1} \frac{\sqrt{n}}{2} B_{\mu,\epsilon}.
\end{aligned} \tag{101}$$

The above inequality enables us to remove  $\text{trace}[\dot{\Sigma}_X]$  from the first line of (99):

$$\begin{aligned}
& \left\| \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \left( 1 + \frac{n-2}{N_{X,\min,\epsilon}(n+2)} \right) \dot{\Sigma}_X \right. \\
& \quad \left. - \frac{n}{N_{X,\min,\epsilon}(n+2)} \cdot \left( 1 + \frac{n^2+n-2}{N_{X,\min,\epsilon}(n+2)} \right)^{-1} \mathbb{E}_{Y_{X,\epsilon}|N_{X,\epsilon},X} [\text{trace} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}]] \cdot I_n \right\|_F \\
&\leq \frac{1}{2} B_{\mu,\epsilon} + \frac{n}{N_{X,\min,\epsilon}(n+2)} \left( 1 + \frac{n^2+n-2}{N_{X,\min,\epsilon}(n+2)} \right)^{-1} \frac{\sqrt{n}}{2} B_{\mu,\epsilon} \cdot \|I_n\|_F \quad (\text{see (99) and (101)}) \\
&= \frac{1}{2} \left( 1 + \frac{n^2}{N_{X,\min,\epsilon}(n+2)} \left( 1 + \frac{n^2+n-2}{N_{X,\min,\epsilon}(n+2)} \right)^{-1} \right) B_{\mu,\epsilon}. \quad (\|I_n\|_F = \sqrt{n})
\end{aligned} \tag{102}$$

Lastly, (102) can be rewritten as follows by introducing  $\ddot{\Sigma}_{X,Y_{X,\epsilon}} \in \mathbb{R}^{n \times n}$ :

$$\begin{aligned}
& \left\| \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,N_{X,\epsilon},X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \dot{\Sigma}_X \right\|_F \\
&\leq \frac{1}{2} \left( 1 + \frac{n-2}{N_{X,\min,\epsilon}(n+2)} \right)^{-1} \left( 1 + \frac{n^2}{N_{X,\min,\epsilon}(n+2)} \left( 1 + \frac{n^2+n-2}{N_{X,\min,\epsilon}(n+2)} \right)^{-1} \right) B_{\mu,\epsilon} \\
&\leq B_{\mu,\epsilon}. \quad (\text{the factor in front of } B_{\mu,\epsilon} \text{ does not exceed } 1)
\end{aligned} \tag{103}$$

Above, we set

$$\begin{aligned}
& \ddot{\Sigma}_{X,Y_{X,\epsilon}} \\
&:= \left(1 + \frac{n-2}{N_{X,\min,\epsilon}(n+2)}\right)^{-1} \left( \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \frac{n}{N_{X,\min,\epsilon}(n+2)} \left(1 + \frac{n^2+n-2}{N_{X,\min,\epsilon}(n+2)}\right)^{-1} \cdot \text{trace} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] \cdot I_n \right) \\
&= \left(1 + \frac{1 - \frac{2}{n}}{1 + \frac{2}{n}} \cdot N_{X,\min,\epsilon}^{-1}\right)^{-1} \left( \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \left( \left(1 + \frac{2}{n}\right) N_{X,\min,\epsilon} + n + 1 - \frac{2}{n} \right)^{-1} \text{trace} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] \cdot I_n \right) \\
&= \left(1 + \frac{1 - \frac{2}{n}}{1 + \frac{2}{n}} \cdot N_{X,\min,\epsilon}^{-1}\right)^{-1} \\
&\quad \cdot \left( \frac{1}{N} \sum_{x \in X} \ddot{\nabla}_{Y_{X,\epsilon}} f(x) \ddot{\nabla}_{Y_{X,\epsilon}} f(x)^* - \left( \left(1 + \frac{2}{n}\right) N_{X,\min,\epsilon} + n + 1 - \frac{2}{n} \right)^{-1} \frac{1}{N} \sum_{x \in X} \left\| \ddot{\nabla}_{Y_{X,\epsilon}} f(x) \right\|_2^2 \cdot I_n \right) \\
&= \frac{1}{N} \left(1 + \frac{1 - \frac{2}{n}}{1 + \frac{2}{n}} \cdot N_{X,\min,\epsilon}^{-1}\right)^{-1} \cdot \left( \sum_{N_{x,\epsilon} \geq N_{X,\min,\epsilon}} \ddot{\nabla}_{Y_{X,\epsilon}} f(x) \ddot{\nabla}_{Y_{X,\epsilon}} f(x)^* \right. \\
&\quad \left. - \left( \left(1 + \frac{2}{n}\right) N_{X,\min,\epsilon} + n + 1 - \frac{2}{n} \right)^{-1} \sum_{N_{x,\epsilon} \geq N_{X,\min,\epsilon}} \left\| \ddot{\nabla}_{Y_{X,\epsilon}} f(x) \right\|_2^2 \cdot I_n \right), \tag{104}
\end{aligned}$$

where the third identity uses (93) and the last line above follows from (40). Because  $B_{\mu,\epsilon}$  does not depend on  $N_{X,\epsilon}$ , it is easy to remove the conditioning on  $N_{X,\epsilon}$  in (103):

$$\begin{aligned}
\left\| \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \dot{\Sigma}_X \right\|_F &= \left\| \mathbb{E} \left[ \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,N_{X,\epsilon},X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \dot{\Sigma}_X \right] \right\|_F \\
&\leq \mathbb{E} \left\| \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,N_{X,\epsilon},X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \dot{\Sigma}_X \right\|_F \quad (\text{Jensen's inequality}) \\
&\leq \mathbb{E} B_{\mu,\epsilon} \\
&= B_{\mu,\epsilon}. \quad (\text{see (99)}) \tag{105}
\end{aligned}$$

Consider also the following special case. Let  $\mu$  be the uniform probability measure on  $\mathbb{D}$  and fix  $x$  within the  $\epsilon$ -interior of  $\mathbb{D}$ , namely  $x \in \mathbb{D}_\epsilon$ . Also draw  $y$  from  $\mu_{x,\epsilon}$ , namely  $y|x \sim \mu_{x,\epsilon}$  (see (17)). Then, as stated in Proposition 3,  $B'_{\mu,\epsilon} = 0$ . Furthermore, it is known [42] that

$$P_{x,y} \cdot \nabla f(x) \stackrel{\text{dist.}}{=} \omega \cdot \nabla f(x) + \sqrt{\omega - \omega^2} \|\nabla f(x)\|_2 \cdot A\alpha, \tag{106}$$

where  $\omega$  follows the beta distribution,  $\alpha$  is uniformly distributed on the unit sphere in  $\mathbb{R}^{n-1}$ , and the two variables are independent, i.e.,

$$\omega \sim \text{beta} \left( \frac{1}{2}, \frac{n-1}{2} \right), \quad \alpha \sim \text{uniform}(\mathbb{S}^{n-2}), \quad \omega \perp \alpha.$$

Finally,  $A \in \mathbb{R}^{n \times (n-1)}$  in (106) is an orthonormal basis for the directions orthogonal to  $\nabla f(x) \in \mathbb{R}^n$ , namely

$$A^* \nabla f(x) = 0, \quad A^* A = I_{n-1}. \tag{107}$$

Using the expressions for first and second moments of beta distribution in the fourth line below, we write

that

$$\begin{aligned}
& \mathbb{E}_{y|x} [P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}] \\
&= \mathbb{E} \left[ \left( \omega \nabla f(x) + \sqrt{\omega - \omega^2} \|\nabla f(x)\|_2 \cdot A\alpha \right) \cdot \left( \omega \nabla f(x) + \sqrt{\omega - \omega^2} \|\nabla f(x)\|_2 \cdot A\alpha \right)^* \right] \quad (\text{see (106)}) \\
&= \mathbb{E} [\omega^2] \cdot \nabla f(x) \nabla f(x)^* + \mathbb{E} [\omega - \omega^2] \|\nabla f(x)\|_2^2 \cdot A \cdot \mathbb{E} [\alpha \alpha^*] \cdot A^* \quad (\omega \perp \alpha, \quad \mathbb{E} \alpha = 0) \\
&= \frac{3}{n(n+2)} \cdot \nabla f(x) \nabla f(x)^* + \frac{n-1}{n(n+2)} \cdot \|\nabla f(x)\|_2^2 \cdot A \cdot \frac{I_{n-1}}{n-1} \cdot A^* \quad \left( \mathbb{E} [\alpha \alpha^*] = \frac{I_{n-1}}{n-1} \right) \\
&= \frac{3}{n(n+2)} \cdot \nabla f(x) \nabla f(x)^* + \frac{1}{n(n+2)} \cdot \|\nabla f(x)\|_2^2 \cdot A A^* \\
&= \frac{2}{n(n+2)} \cdot \nabla f(x) \nabla f(x)^* + \frac{1}{n(n+2)} \cdot \|\nabla f(x)\|_2^2 \cdot \left( \frac{\nabla f(x)}{\|\nabla f(x)\|_2} \cdot \frac{\nabla f(x)^*}{\|\nabla f(x)\|_2} + A A^* \right) \\
&= \frac{2}{n(n+2)} \cdot \nabla f(x) \nabla f(x)^* + \frac{1}{n(n+2)} \cdot \|\nabla f(x)\|_2^2 \cdot I_n, \quad (\text{see (107)}) \tag{108}
\end{aligned}$$

and, consequently,  $B''_{\mu,\epsilon} = 0$ . Furthermore, assume that  $N_{x,\epsilon} = N_{x',\epsilon}$  for every pair  $x, x' \in X$ . Then, we observe that the upper bound in (99) can be improved to  $B_{\mu,\epsilon} = 0$ , namely  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  is an unbiased estimator of  $\dot{\Sigma}_X$ , conditioned on the event  $\mathcal{E}_2$ . This completes the proof of Lemma 3 after dividing all sides of (105) by  $\sqrt{n}$  and redefining  $B''_{\mu,\epsilon}$  and  $B_{\mu,\epsilon}$  accordingly.

## H Proof of Lemma 4

Throughout,  $X$  is fixed and we assume that the event  $\mathcal{E}_2$  holds (see (40)). To bound the estimation error, we write that

$$\begin{aligned}
& \frac{1}{\sqrt{n}} \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \dot{\Sigma}_X \right\|_F \\
&\leq \frac{1}{\sqrt{n}} \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] \right\|_F + \frac{1}{\sqrt{n}} \left\| \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \dot{\Sigma}_X \right\|_F \quad (\text{triangle inequality}) \\
&\leq \frac{1}{\sqrt{n}} \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] \right\|_F + B_{\mu,X,Y_{X,\epsilon}}. \quad (\text{see Lemma 3}) \tag{109}
\end{aligned}$$

It therefore suffices to study the concentration of  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  about its expectation. In fact, as we show next, it is more convenient to first study the concentration of  $\ddot{\Sigma}_{X,Y_{X,\epsilon}} \in \mathbb{R}^{n \times n}$  instead, where

$$\ddot{\Sigma}_{X,Y_{X,\epsilon}} := \frac{1}{N} \sum_{x \in X} \ddot{V}_{Y_{X,\epsilon}} f(x) \cdot \ddot{V}_{Y_{X,\epsilon}} f(x)^*, \tag{110}$$

$$\mathbb{R}^n \ni \ddot{V}_{Y_{X,\epsilon}} f(x) := \frac{n}{N_{x,\epsilon}} \sum_{y \in Y_{x,\epsilon}} P_{x,y} \cdot \nabla f(x), \quad \forall x \in X.$$

Above,  $Y_{X,\epsilon} = \{Y_{x,\epsilon}\}_x$  and  $\overline{N}_{X,\epsilon} = \{N_{x,\epsilon}\}_x$  (see (16)) capture the neighborhood structure of data. Indeed, conditioned on  $\mathcal{E}_2$ , the expression for  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  in (31) simplifies to

$$\ddot{\Sigma}_{X,Y_{X,\epsilon}} = \left( 1 + \frac{1 - \frac{2}{n}}{1 + \frac{2}{n}} \cdot N_{X,\min,\epsilon}^{-1} \right)^{-1} \left( \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \frac{\text{trace} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}]}{(1 + \frac{2}{n}) N_{X,\min,\epsilon} + n + 1 - \frac{2}{n}} \cdot I_n \right). \tag{111}$$

Consequently, deviation of  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  about its expectation can be bounded as:

$$\begin{aligned}
& \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] \right\|_F \\
& \leq \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] \right\|_F \\
& \quad + \left( \left( 1 + \frac{2}{n} \right) N_{X,\min,\epsilon} + n + 1 - \frac{2}{n} \right)^{-1} \cdot \left| \text{trace} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\text{trace} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}]] \right| \cdot \|I_n\|_F \\
& \leq \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] \right\|_F \\
& \quad + \left( \left( 1 + \frac{2}{n} \right) N_{X,\min,\epsilon} + n + 1 - \frac{2}{n} \right)^{-1} \cdot \sqrt{n} \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] \right\|_F \cdot \sqrt{n} \quad (\|I_n\|_F = \sqrt{n}) \\
& = \left( 1 + \frac{n}{\left( 1 + \frac{2}{n} \right) N_{X,\min,\epsilon} + n + 1 - \frac{2}{n}} \right) \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] \right\|_F \\
& \leq 2 \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] \right\|_F. \quad (\text{the factor above does not exceed } 2) \tag{112}
\end{aligned}$$

Above, the first inequality uses (111). We also used the linearity of trace and the inequality  $|\text{trace}[A]| \leq \sqrt{n}\|A\|_F$  for arbitrary  $A \in \mathbb{R}^{n \times n}$ . Thanks to (112), it suffices to study the concentration of  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  about its expectation. The following result is proved in Appendix I.

**Lemma 5.** Fix  $X$  and  $\epsilon \in (0, \epsilon_{\mu,X}]$ . Conditioned on the event  $\mathcal{E}_2$ , it holds that

$$\frac{1}{\sqrt{n}} \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] \right\|_F \lesssim \gamma_7 \gamma_2 K_\mu L_f^2 \log^3(nN_{X,\epsilon}) \sqrt{\frac{n}{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}}, \tag{113}$$

except with a probability of at most

$$(nN_{X,\epsilon})^{(1-K_\mu^2 \gamma_2) \log(nN_{X,\epsilon})} + n^{2-\log \gamma_7},$$

and provided that

$$N_{X,\min,\epsilon} \gtrsim \gamma_2^2 K_\mu^2 \log^4(nN_{X,\epsilon}).$$

Combining (112) and Lemma 5 completes the proof of Lemma 4.

## I Proof of Lemma 5

Throughout,  $X$  is fixed and the event  $\mathcal{E}_2$  holds. We will also use  $\bar{N}_{X,\epsilon} = \{N_{x,\epsilon}\}_{x \in X}$  to summarize the neighborhood structure of data (see (16)). As in Appendix G, we again decompose  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}$  into “diagonal” and “off-diagonal” components:

$$\begin{aligned}
\ddot{\Sigma}_{X,Y_{X,\epsilon}} &= \frac{1}{N} \sum_{x \in X} \dot{\nabla}_{Y_{x,\epsilon}} f(x) \cdot \dot{\nabla}_{Y_{x,\epsilon}} f(x)^* \quad (\text{see (110)}) \\
&= \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y, y' \in Y_{x,\epsilon}} P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y'} \quad (\text{see (32)}) \\
&= \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y \in Y_{x,\epsilon}} P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y} + \frac{n^2}{N} \sum_{x \in X} \frac{1}{N_{x,\epsilon}^2} \sum_{y, y' \in Y_{x,\epsilon}} 1_{y \neq y'} \cdot P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y'} \\
&=: \ddot{\Sigma}_{X,Y_{X,\epsilon}}^d + \ddot{\Sigma}_{X,Y_{X,\epsilon}}^o. \tag{114}
\end{aligned}$$

This decomposition, in turn, allows us to break down the error into the contribution of the diagonal and off-diagonal components:

$$\begin{aligned}
& \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}] \right\|_F \\
& \leq \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}}^d - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}^d] \right\|_F + \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}}^o - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\ddot{\Sigma}_{X,Y_{X,\epsilon}}^o] \right\|_F. \tag{115}
\end{aligned}$$

We bound the norms on the right-hand side above separately in Appendices J and K, respectively, and report the results below.



**Lemma 6.** Fix  $X$  and  $\epsilon \in (0, \epsilon_{\mu, X}]$ . Consider the event

$$\mathcal{E}_1 := \left\{ Y_{X, \epsilon} \mid \max_{x \in X} \max_{y \in Y_{X, \epsilon}} \|P_{x, y} \nabla f(x)\|_2^2 \leq \frac{Q_{X, \epsilon} L_f^2}{n} \right\}, \quad (116)$$

for  $Q_{X, \epsilon} > 0$  to be set later. Then, conditioned on the event  $\mathcal{E}_2$ , it holds that

$$\left\| \ddot{\Sigma}_{X, Y_{X, \epsilon}}^d - \mathbb{E}_{Y_{X, \epsilon} | \mathcal{E}_2, X} \left[ \ddot{\Sigma}_{X, Y_{X, \epsilon}}^d \right] \right\|_F \lesssim \gamma_6 \cdot \frac{Q_{X, \epsilon} L_f^2 n}{\sqrt{\rho_{\mu, X, \epsilon} N_{X, \epsilon}}}, \quad (117)$$

for  $\gamma_6 \geq 1$  and except with a probability of at most  $e^{-\gamma_6} + \Pr_{Y_{X, \epsilon} | \mathcal{E}_2, X}[\mathcal{E}_1^C]$ .

**Lemma 7.** Fix  $X$  and  $\epsilon \in (0, \epsilon_{\mu, X}]$ . Let  $\tilde{Y}_{X, \epsilon}$  contain  $Y_{X, \epsilon}$  and three independent copies of it. That is,  $\tilde{Y}_{X, \epsilon} = \cup_{x \in X} \tilde{Y}_{x, \epsilon}$ , where each  $\tilde{Y}_{x, \epsilon}$  contains  $Y_{x, \epsilon}$  and three independent copies of it. Consider the event

$$\mathcal{E}_3 := \left\{ \tilde{Y}_{X, \epsilon} \mid \max_{x \in X} \max_{y \in Y_{x, \epsilon}} \max_{i \in [1: n]} \|P_{x, y} e_i\|_2^2 \leq \frac{Q_{X, \epsilon}}{n} \right\} \\ \cap \left\{ \tilde{Y}_{X, \epsilon} \mid \max_{x \in X} \max_{y \in \tilde{Y}_{x, \epsilon}} \|P_{x, y} \nabla f(x)\|_2^2 \leq \frac{Q_{X, \epsilon} L_f^2}{n} \right\}, \quad (118)$$

for  $Q_{X, \epsilon} > 0$  to be set later. Here,  $e_i \in \mathbb{R}^n$  is the  $i$ th canonical vector. Assume that

$$\Pr_{\tilde{Y}_{X, \epsilon} | \bar{N}_{X, \epsilon}, \mathcal{E}_2, X} [\mathcal{E}_3^C] \lesssim \left( \frac{\log n}{N_{X, \min, \epsilon} \rho_{\mu, X, \epsilon} N_{X, \epsilon}} \right)^{\frac{\log n}{2}}, \quad (119)$$

and  $N \geq \log n$ . Then, conditioned on the event  $\mathcal{E}_2$ , it holds that

$$\left\| \ddot{\Sigma}_{X, Y_{X, \epsilon}}^o - \mathbb{E}_{Y_{X, \epsilon} | \mathcal{E}_2, X} \left[ \ddot{\Sigma}_{X, Y_{X, \epsilon}}^o \right] \right\|_F \lesssim \gamma_7 \sqrt{\log n} \cdot \frac{Q_{X, \epsilon}^2 L_f^2 n}{\sqrt{N_{X, \min, \epsilon} \rho_{\mu, X, \epsilon} N_{X, \epsilon}}}, \quad (120)$$

for  $\gamma_7 \geq 1$  and except with a probability of at most  $n^{2-\log \gamma_7}$ .

Before we can apply Lemmas 6 and 7 to the right-hand side of (115), however, we must show that the event  $\mathcal{E}_1 \cap \mathcal{E}_3$  is very likely to happen. Owing to Assumption 1, this is indeed the case for the right choice of  $Q_{X, \epsilon}$  as shown in Appendix L and summarized below.

**Lemma 8.** Fix  $\epsilon \in (0, \epsilon_{\mu}]$  and  $X$ . Suppose that  $Q_{X, \epsilon} = \gamma_2 K_{\mu} \log^2(n N_{X, \epsilon})$  for  $\gamma_2 \geq 1$ . Then, for any  $\bar{N}_{X, \epsilon}$ , it holds that

$$\Pr_{\tilde{Y}_{X, \epsilon} | \bar{N}_{X, \epsilon}, \mathcal{E}_2, X} [\mathcal{E}_1 \cap \mathcal{E}_3] \geq \Pr_{\tilde{Y}_{X, \epsilon} | \mathcal{E}_2, X} [\mathcal{E}_1 \cap \mathcal{E}_3] \geq 1 - (n N_{X, \epsilon})^{(1-K_{\mu}^2 \gamma_2) \log(n N_{X, \epsilon})}.$$

In particular, by applying Lemma 8, we observe that (119) holds true when  $\gamma_2 \gtrsim K_{\mu}^{-2}$ . Revisiting (115), we put all the pieces together to conclude that

$$\left\| \ddot{\Sigma}_{X, Y_{X, \epsilon}} - \mathbb{E}_{Y_{X, \epsilon} | \mathcal{E}_2, X} \left[ \ddot{\Sigma}_{X, Y_{X, \epsilon}} \right] \right\|_F \\ \leq \left\| \ddot{\Sigma}_{X, Y_{X, \epsilon}}^d - \mathbb{E}_{Y_{X, \epsilon} | \mathcal{E}_2, X} \left[ \ddot{\Sigma}_{X, Y_{X, \epsilon}}^d \right] \right\|_F + \left\| \ddot{\Sigma}_{X, Y_{X, \epsilon}}^o - \mathbb{E}_{Y_{X, \epsilon} | \mathcal{E}_2, X} \left[ \ddot{\Sigma}_{X, Y_{X, \epsilon}}^o \right] \right\|_F \quad (\text{see (115)}) \\ \lesssim \gamma_6 \cdot \frac{Q_{X, \epsilon} L_f^2 n}{\sqrt{\rho_{\mu, X, \epsilon} N_{X, \epsilon}}} + \gamma_7 \sqrt{\log n} \cdot \frac{Q_{X, \epsilon}^2 L_f^2 n}{\sqrt{N_{X, \min, \epsilon} \rho_{\mu, X, \epsilon} N_{X, \epsilon}}} \quad (\text{see Lemmas 6 and 7}) \\ \leq \gamma_6 \cdot \frac{Q_{X, \epsilon} L_f^2 n}{\sqrt{\rho_{\mu, X, \epsilon} N_{X, \epsilon}}} + \gamma_7 \sqrt{\log n} \cdot \frac{Q_{X, \epsilon} L_f^2 n}{\sqrt{\rho_{\mu, X, \epsilon} N_{X, \epsilon}}} \quad (\text{if } N_{X, \min, \epsilon} \geq Q_{X, \epsilon}^2) \\ \lesssim \gamma_7 \log n \cdot \frac{Q_{X, \epsilon} L_f^2 n}{\sqrt{\rho_{\mu, X, \epsilon} N_{X, \epsilon}}} \quad (\gamma_6 = \gamma_7 \log n) \\ \lesssim \gamma_7 \gamma_2 \log^3(n N_{X, \epsilon}) \cdot \frac{K_{\mu} L_f^2 n}{\sqrt{\rho_{\mu, X, \epsilon} N_{X, \epsilon}}}, \quad (\text{choice of } Q_{X, \epsilon} \text{ in Lemma 8}) \quad (121)$$

except with a probability of at most

$$\begin{aligned}
& e^{-\gamma_6} + \Pr_{Y_{X,\epsilon} | \mathcal{E}_{2,X}} [\mathcal{E}_1^C] + n^{2-\log \gamma_7} \quad (\text{see Lemmas 6 and 7}) \\
& \leq e^{-\gamma_6} + (nN_{X,\epsilon})^{(1-K_\mu^2 \gamma_2) \log(nN_{X,\epsilon})} + n^{2-\log \gamma_7}, \quad (\text{see Lemma 8}) \\
& = n^{-\gamma_7} + (nN_{X,\epsilon})^{(1-K_\mu^2 \gamma_2) \log(nN_{X,\epsilon})} + n^{2-\log \gamma_7} \quad (\text{choice of } \gamma_6 \text{ in (121)}) \\
& \lesssim (nN_{X,\epsilon})^{(1-K_\mu^2 \gamma_2) \log(nN_{X,\epsilon})} + n^{2-\log \gamma_7}, \tag{122}
\end{aligned}$$

and provided that  $\gamma_2 \gtrsim K_\mu^{-2}$  and

$$N_{X,\min,\epsilon} \gtrsim Q_{X,\epsilon}^2 = \gamma_2^2 K_\mu^2 \log^4(nN_{X,\epsilon}). \quad (\text{see (121) and Lemma 8}) \tag{123}$$

This completes the proof of Lemma 5.

## J Proof of Lemma 6

Throughout,  $X$  and the neighborhood structure  $\bar{N}_{X,\epsilon} = \{N_{x,\epsilon}\}_{x \in X}$  (see (16)) are fixed. Moreover, we assume that the event  $\mathcal{E}_2$  holds (see (40)). By definition of  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}^d$  in (114), we observe that

$$\begin{aligned}
& \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon},\epsilon}^d - \mathbb{E}_{Y_{X,\epsilon} | N_{X,\epsilon}, \mathcal{E}_{2,X}} \left[ \ddot{\Sigma}_{X,Y_{X,\epsilon}}^d \right] \right\|_F \\
& = \frac{n^2}{N} \left\| \sum_{x \in X} \sum_{y \in Y_{x,\epsilon}} \frac{1}{N_x^2} (P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y} - \mathbb{E}_{y | N_{x,\epsilon}, \mathcal{E}_{2,x}} [P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}]) \right\|_F \quad (\text{see (114)}) \\
& =: \frac{n^2}{N} \left\| \sum_{x \in X} \sum_{y \in Y_{x,\epsilon}} A_{x,y} \right\|_F, \tag{124}
\end{aligned}$$

where  $\{A_{x,y}\}_{x,y} \subset \mathbb{R}^{n \times n}$  are zero-mean independent random matrices. To bound this sum, we appeal to Proposition 2 by computing the  $b$  and  $\sigma$  parameters below. For arbitrary  $x \in X$  and  $y \in Y_{x,\epsilon}$ , note that

$$\begin{aligned}
& \|A_{x,y}\|_F \\
& = \frac{1}{N_{x,\epsilon}^2} \left\| P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y} - \mathbb{E}_{y | \bar{N}_{x,\epsilon}, \mathcal{E}_{2,x}} [P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}] \right\|_F \quad (\text{see (124)}) \\
& \leq \frac{1}{N_{x,\epsilon}^2} \|P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}\|_F + \frac{1}{N_{x,\epsilon}^2} \cdot \mathbb{E}_{y | \bar{N}_{x,\epsilon}, \mathcal{E}_{2,x}} \|P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}\|_F \quad (\text{Jensen's inequality}) \\
& = \frac{1}{N_{x,\epsilon}^2} \|P_{x,y} \nabla f(x)\|_2^2 + \frac{1}{N_{x,\epsilon}^2} \cdot \mathbb{E}_{y | \bar{N}_{x,\epsilon}, \mathcal{E}_{2,x}} \|P_{x,y} \nabla f(x)\|_2^2 \\
& \leq \frac{1}{\min_{x \in X} N_{x,\epsilon}^2} \cdot \max_{x \in X} \max_{y \in Y_{x,\epsilon}} \|P_{x,y} \nabla f(x)\|_2^2 + \frac{1}{\min_{x \in X} N_{x,\epsilon}^2} \cdot \max_{x \in X} \frac{K_\mu \|\nabla f(x)\|_2^2}{n} \quad (\text{see (19)}) \\
& \leq \frac{1}{\min_{x \in X} N_{x,\epsilon}^2} \cdot \max_{x \in X} \max_{y \in Y_{x,\epsilon}} \|P_{x,y} \nabla f(x)\|_2^2 + \frac{1}{\min_{x \in X} N_{x,\epsilon}^2} \cdot \frac{K_\mu L_f^2}{n} \quad (\text{see (8)}) \\
& =: \frac{1}{\min_{x \in X} N_{x,\epsilon}^2} \cdot \frac{Q_{X,\epsilon} L_f^2}{n} + \frac{1}{N_{X,\min,\epsilon}^2} \cdot \frac{K_\mu L_f^2}{n} \\
& \lesssim \frac{1}{\min_{x \in X} N_{x,\epsilon}^2} \cdot \frac{Q_{X,\epsilon} L_f^2}{n} \quad (\text{when } Q_{X,\epsilon} \geq K_\mu) \\
& =: b. \tag{125}
\end{aligned}$$

where, for  $Q_{X,\epsilon} \geq K_\mu$  to be set later, we defined the following event:

$$\mathcal{E}_1 := \left\{ Y_{X,\epsilon} \mid \max_{x \in X} \max_{y \in Y_{x,\epsilon}} \|P_{x,y} \nabla f(x)\|_2^2 \leq \frac{Q_{X,\epsilon} L_f^2}{n} \right\}. \tag{126}$$

On the other hand, note that

$$\begin{aligned}
& \sum_{x \in X} \sum_{y \in Y_{x,\epsilon}} \mathbb{E}_{y|\bar{N}_{x,\epsilon}, \mathcal{E}_{2,x}} \|A_{x,y}\|_F^2 \\
&= \sum_{x \in X} \sum_{y \in Y_{x,\epsilon}} \frac{1}{N_{x,\epsilon}^2} \cdot \mathbb{E}_{y|\bar{N}_{x,\epsilon}, \mathcal{E}_{2,x}} \left\| P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y} - \mathbb{E}_{y|\bar{N}_{x,\epsilon}, \mathcal{E}_{2,x}} [P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}] \right\|_F^2 \\
&\leq \sum_{x \in X} \sum_{y \in Y_{x,\epsilon}} \frac{1}{N_{x,\epsilon}^2} \cdot \mathbb{E}_{y|\bar{N}_{x,\epsilon}, \mathcal{E}_{2,x}} \|P_{x,y} \nabla f(x) \nabla f(x)^* P_{x,y}\|_F^2 \\
&= \sum_{x \in X} \sum_{y \in Y_{x,\epsilon}} \frac{1}{N_{x,\epsilon}^2} \cdot \mathbb{E}_{y|\bar{N}_{x,\epsilon}, \mathcal{E}_{2,x}} \|P_{x,y} \nabla f(x)\|_2^4 \\
&\lesssim \sum_{x \in X} \sum_{y \in Y_{x,\epsilon}} \frac{1}{N_{x,\epsilon}^2} \cdot \frac{K_\mu^2}{n^2} \cdot \max_{x \in X} \|\nabla f(x)\|_2^4 \quad (\text{see (19)}) \\
&\leq \frac{N}{\min_{x \in X} N_{x,\epsilon}} \cdot \frac{K_\mu^2 L_f^4}{n^2}. \quad (\#X = N, \#Y_{x,\epsilon} = N_{x,\epsilon}, \text{ see (8)})
\end{aligned} \tag{127}$$

where the second line uses (124). The third line above uses the fact that  $\mathbb{E} \|Z - \mathbb{E}[Z]\|_F^2 \leq \mathbb{E} \|Z\|_F^2$  for a random matrix  $Z$ . It follows that

$$\max[b, \sigma] \leq \sqrt{\frac{N}{\min_{x \in X} N_{x,\epsilon}}} \cdot \frac{Q_{X,\epsilon} L_f^2}{n}, \quad \text{if } Q_{X,\epsilon} \geq K_\mu. \tag{128}$$

In light of Proposition 2, and conditioned on the event  $\mathcal{E}_1 \cap \mathcal{E}_2$ , it follows that

$$\begin{aligned}
& \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}}^d - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1, \bar{N}_{X,\epsilon}, \mathcal{E}_2, X} \left[ \ddot{\Sigma}_{X,Y_{X,\epsilon}}^d \right] \right\|_F \\
&= \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}}^d - \mathbb{E}_{Y_{X,\epsilon}|\bar{N}_{X,\epsilon}, \mathcal{E}_2, X} \left[ \ddot{\Sigma}_{X,Y_{X,\epsilon}}^d \right] \right\|_F \\
&= \frac{n^2}{N} \left\| \sum_{x \in X} \sum_{y \in Y_{x,\epsilon}} A_{x,y} \right\|_F \quad (\text{see (124)}) \\
&\lesssim \frac{n^2}{N} \cdot \gamma_6 \cdot \max[b, \sigma] \\
&\lesssim \frac{n^2}{N} \cdot \gamma_6 \sqrt{\frac{N}{\min_x N_{x,\epsilon}}} \cdot \frac{Q_{X,\epsilon} L_f^2}{n} \quad (\text{see (128)}) \\
&= \gamma_6 \cdot \frac{n}{\sqrt{N \cdot \min_x N_{x,\epsilon}}} \cdot Q_{X,\epsilon} L_f^2 \\
&\lesssim \gamma_6 \cdot \frac{n}{\sqrt{N \cdot \min_x \frac{\mu(\mathbb{B}_{x,\epsilon})}{\mu(\mathbb{B}_{X,\epsilon})} \cdot N_{X,\epsilon}}} \cdot Q_{X,\epsilon} L_f^2 \quad (\text{see (40)}) \\
&= \gamma_6 \cdot \frac{n}{\sqrt{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \cdot Q_{X,\epsilon} L_f^2, \quad (\text{see (22)})
\end{aligned} \tag{129}$$

for  $\gamma_6 \geq 1$  and except with a probability of at most  $e^{-\gamma_6}$ . The bound in (129) is independent of  $\bar{N}_{x,\epsilon}$ , which we can therefore remove from the picture:

$$\begin{aligned}
& \Pr_{Y_{X,\epsilon}|\mathcal{E}_1, \mathcal{E}_2, X} \left[ \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}}^d - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1, \mathcal{E}_2, X} \left[ \ddot{\Sigma}_{X,Y_{X,\epsilon}}^d \right] \right\|_F \gtrsim \gamma_6 \cdot \frac{Q_{X,\epsilon} L_f^2 n}{\sqrt{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \right] \\
&= \mathbb{E}_{Y_{x,\epsilon}|\mathcal{E}_1, \mathcal{E}_2, X} \left[ \Pr_{Y_{X,\epsilon}|\mathcal{E}_1, \bar{N}_{X,\epsilon}, \mathcal{E}_2, X} \left[ \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}}^d - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1, \bar{N}_{X,\epsilon}, \mathcal{E}_2, X} \left[ \ddot{\Sigma}_{X,Y_{X,\epsilon}}^d \right] \right\|_F \gtrsim \gamma_6 \cdot \frac{Q_{X,\epsilon} L_f^2 n}{\sqrt{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \right] \right] \\
&\leq \mathbb{E}_{Y_{x,\epsilon}|\mathcal{E}_1, \mathcal{E}_2, X} [e^{-\gamma_6}] \quad (\text{see (129)}) \\
&\leq e^{-\gamma_6}.
\end{aligned} \tag{130}$$

Lastly, by applying (57), we remove the conditioning on the event  $\mathcal{E}_1$  as follows:

$$\begin{aligned}
& \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,X} \left[ \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}}^d - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} \left[ \ddot{\Sigma}_{X,Y_{X,\epsilon}}^d \right] \right\|_F \gtrsim \gamma_6 \cdot \frac{Q_{X,\epsilon} L_f^2 n}{\sqrt{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \right] \\
& \leq \Pr_{Y_{X,\epsilon}|\mathcal{E}_1,\mathcal{E}_2,X} \left[ \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}}^d - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} \left[ \ddot{\Sigma}_{X,Y_{X,\epsilon}}^d \right] \right\|_F \gtrsim \gamma_6 \cdot \frac{Q_{X,\epsilon} L_f^2 n}{\sqrt{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \right] + \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\mathcal{E}_1^C] \quad (\text{see (57)}) \\
& = \Pr_{Y_{X,\epsilon}|\mathcal{E}_1,\mathcal{E}_2,X} \left[ \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}}^d - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_1,\mathcal{E}_2,X} \left[ \ddot{\Sigma}_{X,Y_{X,\epsilon}}^d \right] \right\|_F \gtrsim \gamma_6 \cdot \frac{Q_{X,\epsilon} L_f^2 n}{\sqrt{\rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \right] + \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\mathcal{E}_1^C] \\
& \leq e^{-\gamma_6} + \Pr_{Y_{X,\epsilon}|\mathcal{E}_2,X} [\mathcal{E}_1^C]. \quad (\text{see (130)}) \tag{131}
\end{aligned}$$

The proof of Lemma 6 is now complete.

## K Proof of Lemma 7

Throughout,  $X$  and the neighborhood structure  $\overline{N}_{X,\epsilon} = \{N_{x,\epsilon}\}_x$  are fixed (see (16)). We further assume that the event  $\mathcal{E}_2$  holds (see (40)). Let us index  $X$  as  $X = \{x_s\}_{s=1}^N$ . For each  $x_s \in X$ , we index its neighbors  $Y_{x_s,\epsilon}$  as  $Y_{x_s,\epsilon} = \{y_{sk}\}_{k=1}^{N_{x_s,\epsilon}}$ , where  $N_{x_s,\epsilon} = \#Y_{x_s,\epsilon}$  is the number of neighbors of  $x_s$  (within radius of  $\epsilon$ ). Recalling the definition of  $\ddot{\Sigma}_{X,Y_{X,\epsilon}}^o$  from (114), our objective here is to find an upper bound for

$$\begin{aligned}
& \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}}^o - \mathbb{E}_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X} \left[ \ddot{\Sigma}_{X,Y_{X,\epsilon}}^o \right] \right\|_F \\
& = \frac{n^2}{N} \left\| \sum_{s=1}^N \sum_{k \neq l} \frac{1}{N_{x_s,\epsilon}^2} (P_{x_s,y_{sk}} \nabla f(x_s) \nabla f(x_s)^* P_{x_s,y_{sl}} - \mathbb{E}_{y_s|x_s} [P_{x_s,y_s}] \nabla f(x) \nabla f(x)^* \mathbb{E}_{y_s|x_s} [P_{x_s,y_s}]) \right\|_F \\
& =: \frac{n^2}{N} \left\| \sum_{s=1}^N \sum_{k,l=1}^{N_{x_s,\epsilon}} A_{skl} \right\|_F \\
& = \frac{n^2}{N} \sqrt{\sum_{i,j=1}^n \left| \sum_{s=1}^N \sum_{k,l=1}^{N_{x_s,\epsilon}} A_{skl}[i,j] \right|^2}. \tag{132}
\end{aligned}$$

where, in the second line,  $y_s|x_s \sim \mu_{x,\epsilon}$ . Above, we also conveniently defined the matrices  $\{A_{skl}\}_{s,k,l} \subset \mathbb{R}^{n \times n}$  as

$$A_{skl} := \frac{1}{N_{x_s,\epsilon}^2} \begin{cases} P_{x_s,y_{sk}} \nabla f(x_s) \nabla f(x_s)^* P_{x_s,y_{sl}} - \mathbb{E}_{y_s|x_s} [P_{x_s,y_s}] \nabla f(x) \nabla f(x)^* \mathbb{E}_{y_s|x_s} [P_{x_s,y_s}], & k \neq l, \\ 0, & k = l, \end{cases} \tag{133}$$

for every  $s \in [1 : N]$  and  $k, l \in [1 : N_{x_s,\epsilon}]$ . By their definition above, the random matrices  $\{A_{skl}\}_{s,k,l}$  enjoy the following properties:

$$A_{skk} = 0, \quad \mathbb{E}_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X} [A_{skl}] = 0, \quad s \in [1 : N], \quad k, l \in [1 : N_{x_s,\epsilon}]. \tag{134}$$

With fixed  $s \in [1 : N]$  and  $i, j \in [1 : n]$ , we may use  $\{A_{skl}[i, j]\}_{k,l}$  to form a new matrix  $A_{sij}$  as

$$A_{sij} := [A_{skl}[i, j]]_{k,l} \in \mathbb{R}^{N_{x_s,\epsilon} \times N_{x_s,\epsilon}},$$

or, equivalently,

$$A_{sij}[k, l] := A_{skl}[i, j], \quad k, l \in [1 : N_{x_s,\epsilon}]. \tag{135}$$

Let  $A_{ij}$  be the block-diagonal matrix formed from  $\{A_{sij}\}_s \subset \mathbb{R}^{N_{x_s,\epsilon} \times N_{x_s,\epsilon}}$ , i.e.,

$$A_{ij} = \begin{bmatrix} A_{1ij} & & & \\ & A_{2ij} & & \\ & & \ddots & \\ & & & A_{N_X ij} \end{bmatrix} \in \mathbb{R}^{N_{X,\epsilon} \times N_{X,\epsilon}}. \tag{136}$$

where we used the fact that  $N_{X,\epsilon} = \sum_{s=1}^N N_{x_s}$  to calculate the dimensions of  $A_{ij}$ . In particular, (134) implies that

$$A_{ij}[sk, sk] = 0, \\ \mathbb{E}_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon}, \mathcal{E}_2, X} [A_{ij}[sk, tl]] = 0, \quad s, t \in [1 : N], \quad k \in [1 : N_{x_s, \epsilon}], \quad l \in [1 : N_{x_t, \epsilon}], \quad (137)$$

where, ignoring the standard convention, we indexed the entries of  $A_{ij}$  so that  $sk$  corresponds to the  $k$ th row of  $s$ th block (and hence does not stand for the product of  $s$  and  $k$ ). With this new notation, we revisit (132) to write that

$$\begin{aligned} \left\| \overset{\circ}{\Sigma}_{X, Y_{X,\epsilon}} - \mathbb{E}_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon}, \mathcal{E}_2, X} \left[ \overset{\circ}{\Sigma}_{X, Y_{X,\epsilon}} \right] \right\|_F &= \frac{n^2}{N} \sqrt{\sum_{i,j=1}^n \left| \sum_{s,t=1}^N \sum_{k=1}^{N_{x_s, \epsilon}} \sum_{l=1}^{N_{x_t, \epsilon}} A_{ij}[sk, tl] \right|^2} \quad (\text{see (132)}) \\ &=: \frac{n^2}{N} \sqrt{\sum_{i,j=1}^n a_{ij}^2}. \end{aligned} \quad (138)$$

For fixed  $i, j \in [1 : n]$ , let us next focus on the random variable  $a_{ij}$ .

## K.1 Tail Bound for $a_{ij}$

Recall that the  $p$ th moment of a random variable  $z$  is defined as  $\mathbb{E}^p[z] := (\mathbb{E}[|z|^p])^{\frac{1}{p}}$ . Fix  $i, j \in [1 : n]$ . In order to bound  $a_{ij}$ , we

- first, control its moments, namely

$$\mathbb{E}_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon}, \mathcal{E}_2, X}^p [a_{ij}] = \mathbb{E}_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon}, \mathcal{E}_2, X}^p \left[ \sum_{s,t=1}^N \sum_{k=1}^{N_{x_s, \epsilon}} \sum_{l=1}^{N_{x_t, \epsilon}} A_{ij}[sk, tl] \right], \quad \forall p \geq 1. \quad (139)$$

- Second, we use the Markov's inequality to find a tail bound for  $a_{ij}$  (given its moments).

Each step is discussed in a separate subsection below.

### K.1.1 Moments of $a_{ij}$

In order to control the moments of  $a_{ij}$ , we take the following steps:

- *symmetrization*,
- *decoupling*,
- *modulation* with *Rademacher sequences*, and finally
- bounding the moments of the resulting *symmetric decoupled chaos* random variable.

Each of these steps is detailed in a separate paragraph below.

**Symmetrization** To control the moments of  $a_{ij}$ , we first use a symmetrization argument as follows. With  $s \in [1 : N]$  and conditioned on  $x_s$ , let  $Y_{x_s, \epsilon}^i \in \mathbb{R}^{n \times N_{x_s, \epsilon}}$  be an independent copy of  $Y_{x_s, \epsilon}$ . Similar to (133), define a family of matrices  $\{A_{skl}^i\}_{skl} \subset \mathbb{R}^{n \times n}$  as

$$A_{skl}^i := \frac{1}{N_{x_s, \epsilon}^2} \begin{cases} P_{x_s, y_{sk}^i} \nabla f(x_s) \nabla f(x_s)^* P_{x_s, y_{sl}^i} - \mathbb{E}_{y_s|x_s} [P_{x_s, y_s}] \nabla f(x) \nabla f(x)^* \mathbb{E}_{y_s|x_s} [P_{x_s, y_s}], & k \neq l, \\ 0, & k = l, \end{cases} \quad (140)$$

for every  $s \in [1 : N]$  and  $k, l \in [1 : N_{x_s, \epsilon}]$ . As before,  $y_s|x_s \sim \mu_{x, \epsilon}$  (see (17)). With  $s, i, j$  all fixed, we use  $\{A_{skl}^i[i, j]\}_{k,l}$  to form a new matrix  $A_{sij}^i$  as

$$A_{sij}^i := [A_{skl}^i[i, j]]_{k,l} \in \mathbb{R}^{N_{x_s, \epsilon} \times N_{x_s, \epsilon}},$$

or, equivalently,

$$A_{sij}^i[k, l] = A_{skl}^i[i, j], \quad \forall k, l \in [1 : N_{x_s, \epsilon}], \quad (141)$$

as we did in (135). Also, as in (136), form the block-diagonal matrix  $A_{ij}^i \in \mathbb{R}^{N_{X, \epsilon} \times N_{X, \epsilon}}$ . In summary, each  $A_{ij}^i[sk, tl]$  is an independent copy of  $A_{ij}[sk, sl]$  and, similar to (137), we have that

$$A_{ij}^i[sk, sk] = 0, \quad \mathbb{E}_{Y_{X, \epsilon}^i | \overline{N}_{X, \epsilon}, \mathcal{E}_2, X} [A_{ij}[sk, tl]] = 0, \quad s, t \in [1 : N], \quad k \in [1 : N_{x_s, \epsilon}], \quad l \in [1 : N_{x_l, \epsilon}]. \quad (142)$$

Using the above construction, we then write that

$$\begin{aligned} & E_{Y_{X, \epsilon}^i | \overline{N}_{X, \epsilon}, \mathcal{E}_2, X}^p [a_{ij}] \\ &= \mathbb{E}_{Y_{X, \epsilon}^i | \overline{N}_{X, \epsilon}, \mathcal{E}_2, X}^p \left[ \sum_{s,t=1}^N \sum_{k=1}^{N_{x_s, \epsilon}} \sum_{l=1}^{N_{x_t, \epsilon}} A_{ij}[sk, tl] \right] \quad (\text{see (138)}) \\ &= \mathbb{E}_{Y_{X, \epsilon}^i | \overline{N}_{X, \epsilon}, \mathcal{E}_2, X}^p \left[ \sum_{s,t=1}^N \sum_{k=1}^{N_{x_s, \epsilon}} \sum_{l=1}^{N_{x_t, \epsilon}} A_{ij}[sk, tl] - \mathbb{E}_{Y_{X, \epsilon}^i | \overline{N}_{X, \epsilon}, \mathcal{E}_2, X} [A_{ij}^i[sk, tl]] \right] \quad (\text{see (142)}) \\ &= \mathbb{E}_{Y_{X, \epsilon}^i | \overline{N}_{X, \epsilon}, \mathcal{E}_2, X}^p \left[ \mathbb{E}_{Y_{X, \epsilon}^i | \overline{N}_{X, \epsilon}, X} \left[ \sum_{s,t=1}^N \sum_{k=1}^{N_{x_s, \epsilon}} \sum_{l=1}^{N_{x_t, \epsilon}} A_{ij}[sk, tl] - A_{ij}^i[sk, tl] \right] \right] \quad (\text{independence}) \\ &= \mathbb{E}_{Y_{X, \epsilon}^i, Y_{X, \epsilon}^i | \overline{N}_{X, \epsilon}, \mathcal{E}_2, X}^p \left[ \sum_{s,t=1}^N \sum_{k=1}^{N_{x_s, \epsilon}} \sum_{l=1}^{N_{x_t, \epsilon}} \underbrace{A_{ij}[sk, tl] - A_{ij}^i[sk, tl]}_{B_{ij}[sk, tl]} \right], \quad (\text{Jensen's inequality}) \quad (143) \end{aligned}$$

where we defined the block-diagonal matrix  $B_{ij} \in \mathbb{R}^{N_{X, \epsilon} \times N_{X, \epsilon}}$  such that

$$\begin{aligned} & B_{ij}[sk, tl] \\ &:= A_{ij}[sk, tl] - A_{ij}^i[sk, tl] \\ &= \begin{cases} A_{sij}[k, l] - A_{sij}^i[k, l], & s = t \\ 0, & s \neq t, \end{cases} \quad (\text{see (136)}) \\ &= \begin{cases} A_{skl}[i, j] - A_{skl}^i[i, j], & s = t \\ 0, & s \neq t \end{cases} \quad (\text{see (135) and (141)}) \\ &= \begin{cases} N_{x_s, \epsilon}^{-2} \cdot e_i^* \left( P_{x_s, y_{sk}} \nabla f(x_s) \nabla f(x_s)^* P_{x_s, y_{sl}} - P_{x_s, y_{sk}} \nabla f(x_s) \nabla f(x_s)^* P_{x_s, y_{sl}} \right) e_j, & s = t \text{ and } k \neq l, \\ 0, & s \neq t \text{ or } k = l, \end{cases} \quad (144) \end{aligned}$$

for every  $s, t \in [1 : N]$ ,  $k \in [1 : N_{x_s, \epsilon}]$ ,  $l \in [1 : N_{x_t, \epsilon}]$ . Above, the last line uses (133) and (140). Also,  $e_i \in \mathbb{R}^n$  is the  $i$ th coordinate vector. Note that, by construction, each  $B_{ij}[sk, tl]$  is a *symmetric random variable* (in the sense that its distribution is symmetric about the origin). Moreover, similar to (137), it holds that

$$B_{ij}[sk, sk] = 0, \quad \mathbb{E}_{Y_{X, \epsilon}^i, Y_{X, \epsilon}^i | \overline{N}_{X, \epsilon}, \mathcal{E}_2, X} [B_{ij}[sk, tl]] = 0, \quad s, t \in [1 : N], \quad k \in [1 : N_{x_s, \epsilon}], \quad l \in [1 : N_{x_l, \epsilon}]. \quad (145)$$

Our next step is to decouple the sum in the last line of (143).

**Decoupling** Let  $\Xi = \{\xi_{sk}\}_{s,k}$  (with  $s \in [1 : N]$  and  $k \in [1 : N_{x_s, \epsilon}]$ ) be a sequence of independent standard Bernoulli random variables: each  $\xi_{sk}$  independently takes one and zero with equal probabilities. We will shortly use the following simple observation:

$$\mathbb{E}_{\Xi} [\xi_{sk} (1 - \xi_{tl})] = \frac{1}{4}, \quad sk \neq tl. \quad (146)$$

We now revisit (143) and write that

$$\begin{aligned}
& \mathbb{E}_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}^p [a_{ij}] \\
& \leq \mathbb{E}_{Y_{X,\epsilon},Y_{X,\epsilon}^i|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}^p \left[ \sum_{s,t=1}^N \sum_{k=1}^{N_{x_s,\epsilon}} \sum_{l=1}^{N_{x_t,\epsilon}} B_{ij}[sk,tl] \right] \quad (\text{see (143)}) \\
& = 4 \cdot \mathbb{E}_{Y_{X,\epsilon},Y_{X,\epsilon}^i|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}^p \left[ \sum_{s,t=1}^N \sum_{k=1}^{N_{x_s,\epsilon}} \sum_{l=1}^{N_{x_t,\epsilon}} \mathbb{E}_{\Xi} [\xi_{sk} (1 - \xi_{tl}) \cdot B_{ij}[sk,tl] \right] \quad (B_{ij}[sk,sk] = 0, \text{ and (146)}) \\
& \leq 4 \cdot \mathbb{E}_{Y_{X,\epsilon},Y_{X,\epsilon}^i,\Xi|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}^p \left[ \sum_{s,t=1}^N \sum_{k=1}^{N_{x_s,\epsilon}} \sum_{l=1}^{N_{x_t,\epsilon}} \xi_{sk} (1 - \xi_{tl}) \cdot B_{ij}[sk,tl] \right]. \quad (\text{Jensen's inequality}) \quad (147)
\end{aligned}$$

In particular, there must exist  $\Xi_0 = \{\xi_{0sk}\}_{s,k}$  that exceeds the expectation in the last line above, so that

$$\begin{aligned}
& \mathbb{E}_{Y_{X,\epsilon}|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}^p [a_{ij}] \\
& \leq 4 \cdot \mathbb{E}_{Y_{X,\epsilon},Y_{X,\epsilon}^i|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}^p \left[ \sum_{s,t=1}^N \sum_{k=1}^{N_{x_s,\epsilon}} \sum_{l=1}^{N_{x_t,\epsilon}} \xi_{0sk} (1 - \xi_{0tl}) \cdot B_{ij}[sk,tl] \right] \\
& = 4 \cdot \mathbb{E}_{Y_{X,\epsilon},Y_{X,\epsilon}^i|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}^p \left[ \sum_{\xi_{0sk}=1, \xi_{0tl}=0} B_{ij}[sk,tl] \right] \\
& = 4 \cdot \mathbb{E}_{Y_{X,\epsilon},Y_{X,\epsilon}^i|\overline{N}_{X,\epsilon},\mathcal{E}_2,X}^p \left[ \sum_{sk \in S_0, tl \notin S_0} B_{ij}[sk,tl] \right]. \quad (S_0 := \{sk : \xi_{0sk} = 1\} \subseteq [1 : N_{X,\epsilon}]) \quad (148)
\end{aligned}$$

After setting  $\{Y_{X,\epsilon}^{ii}, Y_{X,\epsilon}^{iii}\} \subset \mathbb{R}^{n \times \#N_{X,\epsilon}}$  to be an independent copy of  $\{Y_{X,\epsilon}, Y_{X,\epsilon}^i\}$ , we define the family of matrices  $\{C_{skl}\}_{s,k,l} \subset \mathbb{R}^{n \times n}$  as

$$C_{skl} := \frac{1}{N_{x_s,\epsilon}^2} \begin{cases} P_{x_s, y_{sk}} \nabla f(x_s) \nabla f(x_s)^* P_{x_s, y_{sl}^{ii}} - P_{x_s, y_{sk}^i} \nabla f(x_s) \nabla f(x_s)^* P_{x_s, y_{sl}^{iii}}, & k \neq l, \\ 0, & k = l, \end{cases} \quad (149)$$

for every  $s \in [1 : N]$  and  $k, l \in [1 : N_{x_s,\epsilon}]$ . For fixed  $s, i, j$ , we use  $\{C_{skl}[i, j]\}_{k,l}$  to form a new matrix  $C_{sij}$  as

$$C_{sij} := [C_{skl}[i, j]]_{k,l} \in \mathbb{R}^{N_{x_s,\epsilon} \times N_{x_s,\epsilon}},$$

or, equivalently,

$$C_{sij}[k, l] = C_{skl}[i, j], \quad \forall k, l \in [1 : N]. \quad (150)$$

As before, we also form the block-diagonal matrix  $C_{ij} \in \mathbb{R}^{N_{X,\epsilon} \times N_{X,\epsilon}}$  from  $\{C_{sij}\}_s$ , and record its basic properties:

$$C_{ij}[sk, sk] = 0,$$

$$\mathbb{E}_{Y_{X,\epsilon}, Y_{X,\epsilon}^i, Y_{X,\epsilon}^{ii}, Y_{X,\epsilon}^{iii}|\overline{N}_{X,\epsilon}, \mathcal{E}_2, X} [C_{ij}[sk, tl]] = 0, \quad s, t \in [1 : N], \quad k \in [1 : N_{x_s,\epsilon}], \quad l \in [1 : N_{x_t,\epsilon}]. \quad (151)$$

For the sake of brevity, we will use the following short hand:

$$\begin{aligned}
\tilde{Y}_{X,\epsilon} &:= Y_{X,\epsilon} \cup Y_{X,\epsilon}^i \cup Y_{X,\epsilon}^{ii} \cup Y_{X,\epsilon}^{iii}, \\
\tilde{Y}_{x_s,\epsilon} &:= Y_{x_s,\epsilon} \cup Y_{x_s,\epsilon}^i \cup Y_{x_s,\epsilon}^{ii} \cup Y_{x_s,\epsilon}^{iii}, \quad \forall x_s \in X. \quad (152)
\end{aligned}$$

Equipped with the construction above, we revisit (148) and write that

$$\begin{aligned}
& \mathbb{E}_{Y_{X,\epsilon}|\bar{N}_{X,\epsilon},\mathcal{E}_2,X}^p [a_{ij}] \\
& \leq 4 \cdot \mathbb{E}_{Y_{X,\epsilon},Y_{X,\epsilon}^i|\bar{N}_{X,\epsilon},\mathcal{E}_2,X}^p \left[ \sum_{sk \in S_0, tl \notin S_0} B_{ij}[sk, tl] \right] \quad (\text{see (148)}) \\
& = 4 \cdot \mathbb{E}_{Y_{X,\epsilon},Y_{X,\epsilon}^i|\bar{N}_{X,\epsilon},X}^p \left[ \sum_{sk \in S_0, sl \notin S_0} N_{x_s,\epsilon}^{-2} \cdot e_i^* \left( P_{x_s,y_{sk}} \nabla f(x_s) \nabla f(x_s)^* P_{x_s,y_{sl}} \right. \right. \\
& \quad \left. \left. - P_{x_s,y_{sk}^i} \nabla f(x_s) \nabla f(x_s)^* P_{x_s,y_{sl}^i} \right) e_j \right] \quad (\text{see (144)}) \\
& = 4 \cdot \mathbb{E}_{\tilde{Y}_{X,\epsilon}|\bar{N}_{X,\epsilon},\mathcal{E}_2,X}^p \\
& \quad \left[ \sum_{sk \in S_0, sl \notin S_0} \underbrace{N_{x_s,\epsilon}^{-2} \cdot e_i^* \left( P_{x_s,y_{sk}} \nabla f(x_s) \nabla f(x_s)^* P_{x_s,y_{sl}^i} - P_{x_s,y_{sk}^i} \nabla f(x_s) \nabla f(x_s)^* P_{x_s,y_{sl}^{ii}} \right) e_j}_{C_{ij}[sk,sl]} \right] \quad (\text{independence}) \\
& = 4 \cdot \mathbb{E}_{\tilde{Y}_{X,\epsilon}|\bar{N}_{X,\epsilon},\mathcal{E}_2,X}^p \left[ \sum_{sk \in S_0, sl \notin S_0} C_{ij}[sk, sl] + \sum_{sk \notin S_0, tl \in S_0} \mathbb{E}_{\tilde{Y}_{X,\epsilon}|\bar{N}_{X,\epsilon},\mathcal{E}_2,X} [C_{ij}[sk, tl]] \right] \quad (\text{see (151)}) \\
& \leq 4 \cdot \mathbb{E}_{\tilde{Y}_{X,\epsilon}|\bar{N}_{X,\epsilon},\mathcal{E}_2,X}^p \left[ \sum_{t,s=1}^N \sum_{k=1}^{N_{x_s,\epsilon}} \sum_{l=1}^{N_{x_t,\epsilon}} C_{ij}[sk, tl] \right]. \quad (\text{independence and Jensen's inequality}) \quad (153)
\end{aligned}$$

Recalling that any matrix can be decomposed into symmetric and skew-symmetric parts, we let that

$$D_{ij} := \frac{C_{ij} + C_{ij}^*}{2} \in \mathbb{R}^{N_{x_s,\epsilon} \times N_{x_t,\epsilon}} \quad (154)$$

denote the symmetric part of  $C_{ij}$ . (Symmetry in matrices should not be confused with symmetry in random variables, which was introduced earlier in this section.) For future reference, we also let  $\{D_{sij}\}_s$  denote the diagonal blocks  $D_{ij}$ , where

$$D_{sij} := \frac{C_{sij} + C_{sij}^*}{2} \in \mathbb{R}^{N_{x_s,\epsilon} \times N_{x_s,\epsilon}}. \quad (\text{see (150)}) \quad (155)$$

It therefore follows that

$$\begin{aligned}
\mathbb{E}_{Y_{X,\epsilon}|\bar{N}_{X,\epsilon},\mathcal{E}_2,X}^p [a_{ij}] & \leq 4 \cdot \mathbb{E}_{\tilde{Y}_{X,\epsilon}|\bar{N}_{X,\epsilon},\mathcal{E}_2,X}^p \left[ \sum_{s,t=1}^N \sum_{k=1}^{N_{x_s,\epsilon}} \sum_{l=1}^{N_{x_t,\epsilon}} C_{ij}[sk, tl] \right] \quad (\text{see (153)}) \\
& = 4 \cdot \mathbb{E}_{\tilde{Y}_{X,\epsilon}|\bar{N}_{X,\epsilon},\mathcal{E}_2,X}^p \left[ \sum_{s,t=1}^N \sum_{k=1}^{N_{x_s,\epsilon}} \sum_{l=1}^{N_{x_t,\epsilon}} D_{ij}[sk, tl] \right]. \quad (156)
\end{aligned}$$

The next step is to modulate the sum in the last line above with a Rademacher sequence.

**Modulation with Rademacher Sequences** Fix  $i, j \in [1 : n]$ , and recall the definitions of  $D_{ij}, C_{ij} \in \mathbb{R}^{N_{x_s,\epsilon} \times N_{x_t,\epsilon}}$  from (154) and (149), respectively. Sum of two symmetric and independent random variables is also symmetric. Therefore, conditioned on  $Y_{X,\epsilon}^{ii}$  and  $Y_{X,\epsilon}^{iii}$ , each  $C_{ij}[sk, tl]$  is a symmetric random variable. Additionally, conditioned on  $Y_{X,\epsilon}^{ii}$  and  $Y_{X,\epsilon}^{iii}$ , each  $D_{ij}[sk, tl]$  is a symmetric random variable too. (The same statement is true when conditioned on  $Y_{X,\epsilon}$  and  $Y_{X,\epsilon}^i$  instead.) We then find that, conditioned on  $Y_{X,\epsilon}^{ii}$  and  $Y_{X,\epsilon}^{iii}$ ,

$$\sum_{s,t=1}^N \sum_{k=1}^{N_{x_s,\epsilon}} \sum_{l=1}^{N_{x_t,\epsilon}} D_{ij}[sk, sl] = \sum_{s=1}^N \sum_{k=1}^{N_{x_s,\epsilon}} \left( \sum_{t=1}^N \sum_{l=1}^{N_{x_t,\epsilon}} D_{ij}[sk, tl] \right), \quad (157)$$



is a sum of independent and symmetric random variables, the distributions of which remain invariant under modulation with a Rademacher sequence.

To be more specific, let  $H = \{\eta_{sk}\}_{s,k}$  (with  $s \in [1 : N]$  and  $k \in [1 : N_{x_s, \epsilon}]$ ) be a Rademacher sequence, that is  $\{\eta_{sk}\}_{s,k}$  are independent Bernoulli random variables taking  $\pm 1$  with equal chances. Also let  $H^i = \{\eta_{sk}^i\}_{s,k}$  be an independent copy of  $H$ . Then, we argue that

$$\begin{aligned}
& \mathbb{E}_{Y_{X, \epsilon}^p | \bar{N}_{X, \epsilon}, \mathcal{E}_2, X}^p [d_{ij}] \\
& \leq 4 \cdot \mathbb{E}_{Y_{X, \epsilon}^p | \bar{N}_{X, \epsilon}, \mathcal{E}_2, X}^p \left[ \sum_{s,t=1}^N \sum_{k=1}^{N_{x_s, \epsilon}} \sum_{l=1}^{N_{x_t, \epsilon}} D_{ij}[sk, tl] \right] \quad (\text{see (153)}) \\
& = 4 \cdot \mathbb{E}_{Y_{X, \epsilon}^p | \bar{N}_{X, \epsilon}, \mathcal{E}_2, X}^p \left[ \mathbb{E}_{Y_{X, \epsilon}, Y_{X, \epsilon}^i | Y_{X, \epsilon}^i, Y_{X, \epsilon}^{iii}, \mathcal{E}_2, X}^p \left[ \sum_{s,k} \left( \sum_{t,l} D_{ij}[sk, tl] \right) \right] \right] \quad (\text{see (152)}) \\
& = 4 \cdot \mathbb{E}_{Y_{X, \epsilon}^p | \bar{N}_{X, \epsilon}, \mathcal{E}_2, X}^p \left[ \mathbb{E}_{Y_{X, \epsilon}, Y_{X, \epsilon}^i, H | Y_{X, \epsilon}^i, Y_{X, \epsilon}^{iii}, X}^p \left[ \sum_{s,k} \eta_{sk} \cdot \left( \sum_{t,l} D_{ij}[sk, tl] \right) \right] \right] \quad (\text{independence and symmetry}) \\
& = 4 \cdot \mathbb{E}_{Y_{X, \epsilon}^p | \bar{N}_{X, \epsilon}, \mathcal{E}_2, X}^p \left[ \mathbb{E}_{Y_{X, \epsilon}, Y_{X, \epsilon}^i, H | \bar{N}_{X, \epsilon}, \mathcal{E}_2, X}^p \left[ \sum_{t,l} \left( \sum_{s,k} \eta_{sk} D_{ij}[sk, tl] \right) \right] \right] \\
& = 4 \cdot \mathbb{E}_{Y_{X, \epsilon}^p | \bar{N}_{X, \epsilon}, \mathcal{E}_2, X}^p \left[ \mathbb{E}_{Y_{X, \epsilon}, Y_{X, \epsilon}^i, H^i | Y_{X, \epsilon}, Y_{X, \epsilon}^i, X}^p \left[ \sum_{t,l} \eta_{tl}^i \cdot \left( \sum_{s,k} \eta_{sk} D_{ij}[sk, tl] \right) \right] \right] \quad (\text{independence and symmetry}) \\
& = 4 \cdot \mathbb{E}_{Y_{X, \epsilon}, H, H^i | \bar{N}_{X, \epsilon}, \mathcal{E}_2, X}^p \left[ \sum_{s,t=1}^N \sum_{k=1}^{N_{x_s, \epsilon}} \sum_{l=1}^{N_{x_t, \epsilon}} \eta_{sk} \eta_{tl}^i \cdot D_{ij}[sk, tl] \right] \\
& =: 4 \cdot \mathbb{E}_{Y_{X, \epsilon}, H, H^i | \bar{N}_{X, \epsilon}, \mathcal{E}_2, X}^p [d_{ij}], \tag{158}
\end{aligned}$$

where we set

$$d_{ij} := \left| \sum_{s,t=1}^N \sum_{k=1}^{N_{x_s, \epsilon}} \sum_{l=1}^{N_{x_t, \epsilon}} \eta_{sk} \eta_{tl}^i \cdot D_{ij}[sk, tl] \right|. \tag{159}$$

Conditioned on everything but  $H$  and  $H^i$ ,  $d_{ij}$  is a *symmetric and decoupled chaos*: symmetric because  $D_{ij} \in \mathbb{R}^{N_{X, \epsilon} \times N_{X, \epsilon}}$  is a symmetric matrix, and decoupled because  $H = \{\eta_{sk}\}_{s,k}$  and  $H^i = \{\eta_{sk}^i\}_{s,k}$  are independent (Rademacher) sequences. The behavior of the moments of a chaos random variable is well-understood.

**Moments of Symmetric Decoupled Chaos** The first moment of  $d_{ij}$ , namely its expectation, can be estimated as follows. First observe that

$$\begin{aligned}
\mathbb{E}_{H, H^i | \bar{N}_{X, \epsilon}, \mathcal{E}_2, X} [d_{ij}] & \leq \sqrt{\mathbb{E}_{H, H^i | \bar{N}_{X, \epsilon}, \mathcal{E}_2, X} [d_{ij}^2]} \quad (\text{Jensen's inequality}) \\
& = \|D_{ij}\|_F \quad (H \text{ and } H^i \text{ are independent Rademacher sequences}) \\
& \leq \|C_{ij}\|_F \quad (\text{see (154)}) \\
& \leq \sqrt{N} \cdot \max_{s \in [1:N]} \|C_{sij}\|_F. \quad (C_{ij} \text{ is block-diagonal}) \tag{160}
\end{aligned}$$

Let us therefore focus on  $\|C_{sij}\|_F$  for fixed  $s \in [1 : N]$ :

$$\begin{aligned}
\|C_{sij}\|_F^2 &= \sum_{k,l=1}^{N_{x_s,\epsilon}} |C_{sij}[k,l]|^2 \\
&= \sum_{k,l=1}^{N_{x_s,\epsilon}} |C_{skl}[i,j]|^2 \quad (\text{see (150)}) \\
&\leq N_{x_s,\epsilon}^2 \cdot \max_{k,l \in [1:N_{x_s,\epsilon}]} |C_{skl}[i,j]|^2 \\
&\leq N_{x_s,\epsilon}^2 \cdot \max_{k,l \in [1:N_{x_s,\epsilon}]} \|C_{skl}\|_\infty^2, \tag{161}
\end{aligned}$$

where  $\|A\|_\infty$  is the largest entry of  $A$  in magnitude. With  $e_i \in \mathbb{R}^n$  denoting the  $i$ th canonical vector, we continue by noting that

$$\begin{aligned}
&\|C_{skl}\|_\infty \\
&= \max_{i,j \in [1:n]} |C_{skl}[i,j]| \\
&= \max_{i,j \in [1:n]} |e_i^* C_{skl} e_j| \\
&= N_{x_s,\epsilon}^{-2} \cdot \max_{i,j \in [1:n]} \left| e_i^* \left( P_{x_s,y_{sk}} \nabla f(x_s) \nabla f(x_s)^* P_{x_s,y_{sl}^i} - P_{x_s,y_{sk}^i} \nabla f(x_s) \nabla f(x_s)^* P_{x_s,y_{sl}^{ii}} \right) e_j \right| \quad (\text{see (149)}) \\
&\leq N_{x_s,\epsilon}^{-2} \cdot \max_{i,j \in [1:n]} \left[ \left| e_i^* P_{x_s,y_{sk}} \nabla f(x_s) \nabla f(x_s)^* P_{x_s,y_{sl}^i} e_j \right| + \left| e_i^* P_{x_s,y_{sk}^i} \nabla f(x_s) \nabla f(x_s)^* P_{x_s,y_{sl}^{ii}} e_j \right| \right] \\
&\leq 2N_{x_s,\epsilon}^{-2} \cdot \max_{s \in [1:N]} \max_{i \in [1:n]} \max_{y_s \in \tilde{Y}_{x_s,\epsilon}} |e_i^* P_{x_s,y_s} \nabla f(x_s)|^2 \quad (\text{see (152)}) \\
&\leq 2N_{x_s,\epsilon}^{-2} \cdot \max_{s \in [1:N]} \max_{i \in [1:n]} \max_{y_s \in \tilde{Y}_{x_s,\epsilon}} \|P_{x_s,y_s} e_i\|_2^2 \cdot \|P_{x_s,y_s} \nabla f(x_s)\|_2^2 \quad (\text{Cauchy-Schwarz's inequality}) \\
&\leq 2N_{x_s,\epsilon}^{-2} \cdot \frac{Q_{X,\epsilon}}{n} \cdot \frac{Q_{X,\epsilon} L_f^2}{n}, \quad (\text{conditioned on the event } \mathcal{E}_3) \tag{162}
\end{aligned}$$

where we defined the event  $\mathcal{E}_3$  as

$$\begin{aligned}
\mathcal{E}_3 &= \left\{ \tilde{Y}_{X,\epsilon} \mid \max_{s \in [1:N]} \max_{i \in [1:n]} \max_{y_s \in \tilde{Y}_{x_s,\epsilon}} \|P_{x_s,y_s} e_i\|_2^2 \leq \frac{Q_{X,\epsilon}}{n} \right\} \\
&\cap \left\{ \tilde{Y}_{X,\epsilon} \mid \max_{s \in [1:N]} \max_{y_s \in \tilde{Y}_{x_s,\epsilon}} \|P_{x_s,y_s} \nabla f(x_s)\|_2^2 \leq \frac{Q_{X,\epsilon} L_f^2}{n} \right\}, \tag{163}
\end{aligned}$$

for  $Q_{X,\epsilon} > 0$  to be set later. For  $p \geq 1$  to be assigned later, we also assume that  $\mathcal{E}_3$  is very likely to happen:

$$\Pr_{\tilde{Y}_{X,\epsilon} | \bar{N}_{X,\epsilon}, \mathcal{E}_2, X} [\mathcal{E}_3^C] \lesssim \left( \frac{p}{N_{X,\min,\epsilon} \rho_{\mu,X,\epsilon} N_{X,\epsilon}} \right)^{\frac{p}{2}}, \quad (\text{see (22)}) \tag{164}$$

We now complete our calculation of the first moment of  $d_{ij}$ :

$$\begin{aligned}
\mathbb{E}_{H,H^i | \mathcal{E}_3, \tilde{Y}_{X,\epsilon}, \mathcal{E}_2, X} [d_{ij}] &\leq \|D_{ij}\|_F \\
&\leq \sqrt{N} \cdot \max_{s \in [1:N]} \|C_{sij}\|_F \quad (\text{see (160)}) \\
&\leq \sqrt{N} \cdot \max_{s \in [1:N]} \max_{k,l \in [1:N_{x_s,\epsilon}]} N_{x_s,\epsilon} \cdot \|C_{skl}\|_\infty \quad (\text{see (161)}) \\
&\leq \sqrt{N} \cdot \max_{s \in [1:N]} N_{x_s,\epsilon} \cdot \frac{2Q_{X,\epsilon}^2 L_f^2}{n^2 N_{x_s,\epsilon}^2} \quad (\text{see (162)}) \\
&\leq \frac{2\sqrt{N} Q_{X,\epsilon}^2 L_f^2}{n^2 \cdot \min_{x \in X} N_{x,\epsilon}}. \tag{165}
\end{aligned}$$

To control the higher order moments of  $d_{ij}$ , we invoke the following result [63, Theorem 7].

**Proposition 4. (Moments of a symmetric decoupled chaos)** *For a symmetric matrix  $D$ , a Rademacher sequence  $H = \{\eta_k\}_k$ , and an independent copy  $H^i = \{\eta_l^i\}_l$ , consider the symmetric decoupled (second-order) chaos*

$$d = \left| \sum_{k,l} \eta_k \eta_l^i \cdot D[k, l] \right|.$$

Then, it holds that

$$\mathbb{E}^p[d - \mathbb{E}[d]] \lesssim p \cdot b + \sqrt{p} \cdot \sigma, \quad \forall p \geq 1, \quad (166)$$

where

$$b := \|D\|, \quad (167)$$

$$\sigma^2 := \mathbb{E}_H \left[ \|D\eta\|_2^2 \right], \quad (168)$$

and  $\eta$  is the vector formed from the Rademacher sequence  $H$ .

We now appeal to Proposition 4 in order to bound the moments of the chaos random variable  $d_{ij}$  in (159) (conditioned on  $X, \tilde{Y}_{X,\epsilon}$  and the event  $\mathcal{E}_2 \cap \mathcal{E}_3$ ). To that end, let  $e_k \in \mathbb{R}^n$  be the  $k$ th canonical vector, and note that

$$\begin{aligned} b &= \|D_{ij}\| \quad (\text{see (167)}) \\ &\leq \|C_{ij}\| \quad (\text{see (154)}) \\ &= \max_{s \in [1:N]} \|C_{sij}\|. \quad (C_{ij} \text{ is block-diagonal}) \end{aligned} \quad (169)$$

Let us then focus on  $\|C_{sij}\|$  for fixed  $s \in [1:N]$ . Observe that

$$\begin{aligned} \|C_{sij}\| &\leq N_{x_s, \epsilon} \cdot \|C_{sij}\|_\infty \quad (\|A\| \leq a \cdot \|A\|_\infty, \forall A \in \mathbb{R}^{a \times a}) \\ &= N_{x_s, \epsilon} \cdot \|C_{skl}\|_\infty \quad (\text{see (150)}) \\ &\leq N_{x_s, \epsilon} \cdot \max_{k, l \in [1:N_{x_s, \epsilon}]} \|C_{skl}\|_\infty \\ &\leq N_{x_s, \epsilon} \cdot \frac{2Q_{X, \epsilon}^2 L_f^2}{n^2 N_{x_s, \epsilon}^2} \quad (\text{see (162)}) \\ &\leq \frac{2Q_{X, \epsilon}^2 L_f^2}{n^2 \cdot \min_{x \in X} N_{x, \epsilon}}. \quad (\text{see (40)}) \end{aligned} \quad (170)$$

In light of (169), it follows that

$$\begin{aligned} b &\leq \max_{s \in [1:N]} \|C_{sij}\| \quad (\text{see (169)}) \\ &\leq \frac{2Q_{X, \epsilon}^2 L_f^2}{n^2 \cdot \min_{x \in X} N_{x, \epsilon}}. \quad (\text{see (170)}) \end{aligned} \quad (171)$$

We argue likewise to find  $\sigma$ :

$$\begin{aligned} \sigma &= \sqrt{\mathbb{E}_H \left[ \|D_{ij}\eta\|_2^2 \right]} \quad (\text{see (168)}) \\ &= \|D_{ij}\|_F \quad (\eta \text{ is a Rademacher sequence}) \\ &\leq \frac{2\sqrt{N}Q_{X, \epsilon}^2 L_f^2}{n^2 \cdot \min_{x \in X} N_{x, \epsilon}}. \quad (\text{see (165)}) \end{aligned} \quad (172)$$

With  $b$  and  $\sigma$  at hand, we now invoke Proposition 4 to write that

$$\begin{aligned}
& \mathbb{E}_{H, H^i | \mathcal{E}_3, \tilde{Y}_{X, \epsilon}, \mathcal{E}_2, X}^p [d_{ij}] \\
&= \mathbb{E}_{H, H^i | \mathcal{E}_3, \tilde{Y}_{X, \epsilon}, \mathcal{E}_2, X}^p \left[ \sum_{s, t=1}^N \sum_{k=1}^{N_{x_s, \epsilon}} \sum_{l=1}^{N_{x_t, \epsilon}} \eta_{sk} \eta_{tl}^i \cdot D_{ij}[sk, tl] \right] \quad (\text{see (159)}) \\
&\leq \mathbb{E}_{H, H^i | \mathcal{E}_3, \tilde{Y}_{X, \epsilon}, \mathcal{E}_2, X}^p \left[ d_{ij} - \mathbb{E}_{H, H^i | \mathcal{E}_3, \tilde{Y}_{X, \epsilon}, \mathcal{E}_2, X} [d_{ij}] \right] + \mathbb{E}_{H, H^i | \mathcal{E}_3, \tilde{Y}_{X, \epsilon}, \mathcal{E}_2, X} [d_{ij}] \quad (\text{triangle inequality}) \\
&\lesssim (p \cdot b + \sqrt{p} \cdot \sigma) + \frac{\sqrt{N} Q_{X, \epsilon}^2 L_f^2}{n^2 \cdot \min_x N_{x, \epsilon}} \quad (\text{see Proposition 4 and (165)}) \\
&\lesssim \left( p \cdot \frac{Q_{X, \epsilon}^2 L_f^2}{n^2 \cdot \min_x N_{x, \epsilon}} + \sqrt{p} \cdot \frac{\sqrt{N} Q_{X, \epsilon}^2 L_f^2}{n^2 \cdot \min_x N_{x, \epsilon}} \right) + \frac{\sqrt{N} Q_{X, \epsilon}^2 L_f^2}{n^2 \cdot \min_x N_{x, \epsilon}} \quad (\text{see (171) and (172)}) \\
&\lesssim \sqrt{p} \cdot \frac{\sqrt{N} Q_{X, \epsilon}^2 L_f^2}{n^2 \cdot \min_x N_{x, \epsilon}} \quad (\text{if } 1 \leq p \leq N) \\
&\lesssim \sqrt{p} \cdot \frac{N Q_{X, \epsilon}^2 L_f^2}{n^2 \sqrt{N_{X, \min, \epsilon}} \cdot \rho_{\mu, X, \epsilon} N_{X, \epsilon}} \quad (\text{see (40) and (22)}) \tag{173}
\end{aligned}$$

Conditioned on  $\bar{N}_{X, \epsilon}$ , the bound above is independent of  $\tilde{Y}_{X, \epsilon}$ , which allows us to remove the conditioning and find that

$$\mathbb{E}_{\tilde{Y}_{X, \epsilon}, H, H^i | \mathcal{E}_3, \bar{N}_{X, \epsilon}, \mathcal{E}_2, X}^p [d_{ij}] \lesssim \sqrt{p} \cdot \frac{N Q_{X, \epsilon}^2 L_f^2}{n^2 \sqrt{N_{X, \min, \epsilon}} \cdot \rho_{\mu, X, \epsilon} N_{X, \epsilon}}. \tag{174}$$

As a useful aside, we also record a uniform bound on  $d_{ij}$  for every  $i, j \in [1 : n]$ :

$$\begin{aligned}
|d_{ij}| &= \left| \sum_{s, t=1}^N \sum_{k=1}^{N_{x_s, \epsilon}} \sum_{l=1}^{N_{x_t, \epsilon}} \eta_{sk} \eta_{tl}^i \cdot D_{ij}[sk, tl] \right| \quad (\text{see (159)}) \\
&\leq \sum_{s, t=1}^N \left| \sum_{k=1}^{N_{x_s, \epsilon}} \sum_{l=1}^{N_{x_t, \epsilon}} \eta_{sk} \eta_{tl}^i \cdot D_{ij}[sk, tl] \right| \quad (\text{triangle inequality}) \\
&= \sum_{s=1}^N \left| \sum_{k=1}^{N_{x_s, \epsilon}} \sum_{l=1}^{N_{x_s, \epsilon}} \eta_{sk} \eta_{sl}^i \cdot D_{sij}[k, l] \right| \quad (D_{ij} \text{ is block-diagonal with blocks } D_{sij} \in \mathbb{R}^{N_{x_s, \epsilon} \times N_{x_s, \epsilon}}; \text{ see (155)}) \\
&\leq \sum_{s=1}^N N_{x_s, \epsilon} \cdot \|D_{sij}\| \quad (H, H^i \text{ are Rademacher sequences}) \\
&\leq \sum_{s=1}^N N_{x_s, \epsilon} \cdot \|C_{sij}\| \quad (\text{see (154)}) \\
&\leq \sum_{s=1}^N N_{x_s, \epsilon}^2 \cdot \|C_{sij}\|_{\infty} \quad (\|A\| \leq a \cdot \|A\|_{\infty}, \forall A \in \mathbb{R}^{a \times a}) \\
&= \sum_{s=1}^N N_{x_s, \epsilon}^2 \cdot \|C_{skl}\|_{\infty} \quad (\text{see (150)}) \\
&\leq \sum_{s=1}^N N_{x_s, \epsilon}^2 \cdot \frac{2Q_{X, \epsilon}^2 L_f^2}{n^2 N_{x_s, \epsilon}^2} \quad (\text{see (162)}) \\
&= \frac{2N Q_{X, \epsilon}^2 L_f^2}{n^2}. \tag{175}
\end{aligned}$$

Putting everything back together, we finally argue that

$$\begin{aligned}
\mathbb{E}_{Y_{X,\epsilon}|\bar{N}_{X,\epsilon},\mathcal{E}_2,X}^p[a_{ij}] &\leq 4 \cdot \mathbb{E}_{\tilde{Y}_{X,\epsilon},H,H^i|\bar{N}_{X,\epsilon},\mathcal{E}_2,X}^p[d_{ij}] \quad (\text{see (158)}) \\
&\leq 4 \cdot \mathbb{E}_{\tilde{Y}_{X,\epsilon},H,H^i|\mathcal{E}_3,\bar{N}_{X,\epsilon},\mathcal{E}_2,X}^p[d_{ij}] + 4 \cdot \sup |d_{ij}| \cdot \left( \Pr_{\tilde{Y}_{X,\epsilon}|\bar{N}_{X,\epsilon},\mathcal{E}_2,X}[\mathcal{E}_3^C] \right)^{\frac{1}{p}} \quad (\text{see (57)}) \\
&\lesssim \sqrt{p} \cdot \frac{NQ_{X,\epsilon}^2 L_f^2}{n^2 \sqrt{N_{X,\min,\epsilon} \rho_{\mu,X,\epsilon} N_{X,\epsilon}}} + \frac{NQ_{X,\epsilon}^2 L_f^2}{n^2} \cdot \sqrt{\frac{p}{N_{X,\min,\epsilon} \rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \quad (\text{see (174), (175), and (164)}) \\
&\lesssim \sqrt{p} \cdot \frac{NQ_{X,\epsilon}^2 L_f^2}{n^2 \sqrt{N_{X,\min,\epsilon} \rho_{\mu,X,\epsilon} N_{X,\epsilon}}}, \quad (176)
\end{aligned}$$

when  $1 \leq p \leq N$  (see (173)). At last, (176) describes the moments of the random variable  $a_{ij}$  for fixed  $i, j$  (and conditioned on  $\bar{N}_{X,\epsilon}, \mathcal{E}_2, X$ ).

### K.1.2 Applying the Markov's Inequality

Given the estimates of the moments of  $a_{ij}$  in (176), we can simply apply the Markov's inequality to translate this information into a tail bound for  $a_{ij}$ . Indeed, for arbitrary  $1 \leq p \leq N$  and  $\gamma_8 > 0$ , it holds that

$$\begin{aligned}
\Pr_{Y_{X,\epsilon}|\bar{N}_{X,\epsilon},\mathcal{E}_2,X} [|a_{ij}| > \gamma_8] &= \Pr_{Y_{X,\epsilon}|\bar{N}_{X,\epsilon},\mathcal{E}_2,X} [|a_{ij}|^p > \gamma_8^p] \\
&\leq \left( \frac{\mathbb{E}_{Y_{X,\epsilon}|\bar{N}_{X,\epsilon},\mathcal{E}_2,X}^p[a_{ij}]}{\gamma_8} \right)^p \quad (\text{Markov's inequality}) \\
&\leq \left( \frac{C_1 \sqrt{p} N Q_{X,\epsilon}^2 L_f^2}{\gamma_8 n^2 \sqrt{N_{X,\min,\epsilon} \rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \right)^p, \quad (\text{see (176)}) \quad (177)
\end{aligned}$$

for an absolute constant  $C_1$ . In particular, the choice of

$$\gamma_8 = C_1 \gamma_7 \cdot \sqrt{\log n} \cdot \frac{NQ_{X,\epsilon}^2 L_f^2}{n^2 \sqrt{N_{X,\min,\epsilon} \rho_{\mu,X,\epsilon} N_{X,\epsilon}}}, \quad p = \max[\log n, 1] \leq N, \quad \gamma_7 \geq 1,$$

yields

$$\begin{aligned}
\Pr_{Y_{X,\epsilon}|\bar{N}_{X,\epsilon},\mathcal{E}_2,X} \left[ |a_{ij}| \gtrsim \gamma_7 \cdot \sqrt{\log n} \cdot \frac{NQ_{X,\epsilon}^2 L_f^2}{n^2 \sqrt{N_{X,\min,\epsilon} \rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \right] \\
\leq \gamma_7^{-\log n} = n^{-\log \gamma_7}. \quad (178)
\end{aligned}$$

With the tail bound of  $a_{ij}$  finally available above (for fixed  $i, j \in [1 : n]$  and conditioned on  $\bar{N}_{X,\epsilon}, \mathcal{E}_2, X$ ), we next quantify how  $\sum_{X,Y_{X,\epsilon}}^o$  concentrates about its expectation.

## K.2 Applying the Union Bound

In light of (178) and by applying the union bound to  $\{a_{ij}\}_{i,j}$ , we arrive at the following statement.

$$\begin{aligned}
\Pr_{Y_{X,\epsilon}|\bar{N}_{X,\epsilon},\mathcal{E}_2,X} \left[ \max_{i,j \in [1:n]} |a_{ij}| \lesssim \gamma_7 \cdot \sqrt{\log n} \cdot \frac{NQ_{X,\epsilon}^2 L_f^2}{n^2 \sqrt{N_{X,\min,\epsilon} \rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \right] \\
\geq 1 - n^2 \cdot n^{-\log \gamma_7}. \quad (\text{union bound and (178)}) \quad (179)
\end{aligned}$$

We can now finally revisit (138) and write that

$$\begin{aligned}
\left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}}^o - \mathbb{E}_{Y_{X,\epsilon}|\bar{N}_{X,\epsilon},\mathcal{E}_2,X} \left[ \ddot{\Sigma}_{X,Y_{X,\epsilon}}^o \right] \right\|_F &= \frac{n^2}{N} \sqrt{\sum_{i,j=1}^n a_{ij}^2} \quad (\text{see (138)}) \\
&\leq \frac{n^3}{N} \cdot \max_{i,j \in [1:n]} |a_{ij}| \\
&\leq \frac{n^3}{N} \cdot C_1 \gamma_7 \cdot \sqrt{\log n} \cdot \frac{NQ_{X,\epsilon}^2 L_f^2}{n^2 \sqrt{N_{X,\min,\epsilon} \rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \quad (\text{see (179)}) \\
&= C_1 \gamma_7 \sqrt{\log n} \cdot \frac{nQ_{X,\epsilon}^2 L_f^2}{\sqrt{N_{X,\min,\epsilon} \rho_{\mu,X,\epsilon} N_{X,\epsilon}}}, \tag{180}
\end{aligned}$$

which holds except with a probability of at most  $n^{2-\log \gamma_7}$ , and under (164) (with  $p = \max[\log n, 1]$ ). The above bound is independent of  $\bar{N}_{X,\epsilon}$ , thereby allowing us to remove the conditioning on it:

$$\begin{aligned}
&\Pr_{Y_{X,\epsilon}|\mathcal{E}_2,X} \left[ \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}}^o - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} \left[ \ddot{\Sigma}_{X,Y_{X,\epsilon}}^o \right] \right\|_F \gtrsim \gamma_7 \sqrt{\log n} \cdot \frac{nQ_{X,\epsilon}^2 L_f^2}{\sqrt{N_{X,\min,\epsilon} \rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \right] \\
&= \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} \left[ \Pr_{Y_{X,\epsilon}|\bar{N}_{X,\epsilon},\mathcal{E}_2,X} \left[ \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}}^o - \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} \left[ \ddot{\Sigma}_{X,Y_{X,\epsilon}}^o \right] \right\|_F \gtrsim \gamma_7 \sqrt{\log n} \cdot \frac{nQ_{X,\epsilon}^2 L_f^2}{\sqrt{N_{X,\min,\epsilon} \rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \right] \right] \\
&= \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} \left[ \Pr_{Y_{X,\epsilon}|\bar{N}_{X,\epsilon},\mathcal{E}_2,X} \left[ \left\| \ddot{\Sigma}_{X,Y_{X,\epsilon}}^o - \mathbb{E}_{Y_{X,\epsilon}|\bar{N}_{X,\epsilon},\mathcal{E}_2,X} \left[ \ddot{\Sigma}_{X,Y_{X,\epsilon}}^o \right] \right\|_F \gtrsim \gamma_7 \sqrt{\log n} \cdot \frac{n\sqrt{N}Q_{X,\epsilon}^2 L_f^2}{\sqrt{N_{X,\min,\epsilon} \rho_{\mu,X,\epsilon} N_{X,\epsilon}}} \right] \right] \\
&\leq \mathbb{E}_{Y_{X,\epsilon}|\mathcal{E}_2,X} \left[ n^{2-\log \gamma_6} \right] \\
&= n^{2-\log \gamma_6}. \tag{181}
\end{aligned}$$

The proof of Lemma 7 is now complete.

## L Proof of Lemma 8

Throughout,  $X$  and the neighborhood structure  $\bar{N}_{X,\epsilon} = \{N_{x,\epsilon}\}_{x \in X}$  are fixed. Therefore, for every  $x \in X$ , the columns of matrices  $Y_{x,\epsilon} \in \mathbb{R}^{n \times N_{x,\epsilon}}$  and  $\tilde{Y}_{x,\epsilon} \in \mathbb{R}^{n \times (4N_{x,\epsilon})}$  are random vectors drawn from the conditional probability measure  $\mu_{x,\epsilon}$  (see (17)). For fixed  $x \in X$  and with  $y \sim \mu_{x,\epsilon}$ , recall from Assumption 1 that

$$\Pr_{y|x} \left[ \|P_{x,y} v\|_2^2 > \frac{\gamma_1}{n} \right] \lesssim e^{-K_\mu \gamma_1},$$

for arbitrary (but fixed)  $v \in \mathbb{R}^n$  with  $\|v\|_2 = 1$  and  $\gamma_1 \geq 1$ . The claim in Lemma 8 readily follows with an application of the union bound: For all possible choices of  $x, y, i$ , it holds that

$$\|P_{x,y} e_i\|_2^2 \leq \frac{\gamma_1}{n}, \quad \|P_{x,y} \nabla f(x)\|_2^2 \leq \frac{\gamma_1 \|\nabla f(x)\|_2^2}{n} \leq \frac{\gamma_1 L_f^2}{n}, \quad (\text{see (8)}) \tag{182}$$

except with a probability of at most

$$O(nN_{X,\epsilon})e^{-K_\mu \gamma_1}.$$

With the choice of  $\gamma_1 = Q_{X,\epsilon} = \gamma_2 K_\mu \log^2(nN_{X,\epsilon})$  for  $\gamma_2 \geq 1$ , we find that the event  $\mathcal{E}_1 \cap \mathcal{E}_3$  happens except with a probability of at most  $(nN_{X,\epsilon})^{(1-K_\mu^2 \gamma_2) \log(nN_{X,\epsilon})}$ . This result is independent of  $\bar{N}_{X,\epsilon}$ , which allows us to remove the conditioning on  $\bar{N}_{X,\epsilon}$  and complete the proof of Lemma 8.